Escuela Politécnica Superior

UAM
Universidad Autónoma de Madrid

19|20

Trabajo fin de grado

Réplica y Agregación de Resultados de un Experimento sobre la Usabilidad de un Chatbot

Gemma Merlo Ballesteros

excelencia Campus Internacional UAM CSIC

www.uam.es

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

Doble Grado en Ingeniería Informática y Matemáticas

TRABAJO FIN DE GRADO

Réplica y Agregación de Resultados de un Experimento sobre la Usabilidad de un Chatbot

Autora: Gemma Merlo Ballesteros
Tutora: Silvia Teresita Acuña Castillo

Julio 2020

Gemma Merlo Ballesteros
Réplica y Agregación de Resultados de un Experimento sobre la Usabilidad de un Chatbot

Gemma Merlo Ballesteros
Calle Francisco Tomás y Valiente Nº 11

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

AGRADECIMIENTOS

A Silvia, mi tutora, siempre disponible para enseñarme y guiarme a lo largo de todo el trabajo.

A Ranci y Andrea, que han estado siempre dispuestas a ayudarme.

A todos los compañeros que participaron de forma voluntaria en mi estudio experimental.

A mi querida familia y a mis grandes amigos, por su cariño y su ánimo en todo momento.

Resumen

Los chatbots son agentes conversacionales basados en mensajes. Estos se encuentran en pleno crecimiento, ya que son utilizados cada vez por un mayor número de usuarios en una gran variedad de ámbitos. Al ser sistemas interactivos, los chatbots requieren de las características de usabilidad. No obstante, hay pocos estudios en la literatura que realicen tanto experimentos como sus réplicas y que lleven a cabo la agregación de los resultados de cada uno de los experimentos verdaderos para consolidar los resultados sobre la usabilidad de los chatbots.

En este trabajo se realiza una segunda réplica (con 48 participantes) de un experimento base (54 participantes) que fue realizado en julio de 2019, así como la primera réplica (con 30 participantes) también llevada a cabo en el año 2019, con el objetivo de agregar los datos de los tres estudios experimentales y asegurar la fiabilidad de los resultados mediante las técnicas empleadas de meta-análisis.

Dichos experimentos evalúan la usabilidad del chatbot SOCIO, que, mediante las redes sociales Twitter y Telegram, interpreta los mensajes de los usuarios para construir modelos o meta-modelos, permitiendo a la Ingeniería del Software realizar esta tarea de manera colaborativa, en relación con la aplicación web Creately, que permite construir, entre otros, diagramas de clases.

El diseño del experimento es de tipo *crossover*, en el que el 50 % de los sujetos experimentales emplean Creately para desarrollar la primera tarea del experimento y SOCIO para la segunda, mientras que el otro 50 % de los sujetos experimentales aplican los tratamientos en el orden contrario. Las tareas consisten en realizar diagramas de clases en equipos de tres integrantes.

Tras la realización del experimento, se ejecuta un análisis gráfico a través de la elaboración de *boxplots* de los datos correspondientes a cada métrica de las características de usabilidad, es decir, la eficacia, la eficiencia y la satisfacción, así como también la calidad de los modelos elaborados. Después, para cada métrica se aplica un modelo lineal mixto. Finalmente se realiza el cálculo del tamaño del efecto producido por el tratamiento.

Tras analizar este experimento se determina que el tamaño muestral es insuficiente, por lo que no se han obtenido diferencias significativas generadas por el tratamiento en ninguna métrica. Sin embargo, la tarea parece generar diferencias en todas las métricas a excepción de la satisfacción y el tiempo de realización.

Mediante la agregación de los resultados se obtiene una muestra de tamaño mucho mayor, aumentando así la evidencia estadística para obtener resultados relevantes. De ahí el empleo de técnicas de meta-análisis, que corresponden a instrumentos estadísticos adecuados para combinar resultados de familias de experimentos. En concreto, se ha utilizado el modelo de regresión lineal del tipo *Individual Participant Data Stratified*.

Finalmente, los resultados a nivel de familia de experimentos reflejan diferencias significativas entre SOCIO y Creately con respecto a la eficiencia, la satisfacción y la calidad. Concretamente, se muestra

mayor nivel de eficiencia (tiempo para elaborar una tarea y nº de mensajes intercambiados de discusión durante la misma), satisfacción y precisión del modelo (porcentaje de elementos correctos en el diagrama elaborado por un equipo, en función de los elementos del diagrama ideal) con SOCIO, mientras que Creately parece mejor en términos de las variables recall (porcentaje de elementos del diagrama ideal que se encuentran en el diagrama realizado por un sujeto experimental) y aciertos (tasa de éxito de cada equipo, en comparación con la solución ideal), ambas métricas de la calidad.

De esta manera, se ha comprobado que los resultados a nivel de familia son mucho más precisos que los resultados del experimento individual, debido al aumento en el tamaño muestral en la agregación.

Palabras clave

Chatbot, Usabilidad, Eficacia, Eficiencia, Satisfacción, Diseño *Crossover*, Modelo Lineal Mixto, Agregación

Abstract

Chatbots are messaging-based dialogue agents. They are growing fast, as they are being used more and more by a greater number of users in a wide variety of areas. Being interactive systems, chatbots require usability characteristics. Nevertheless, there are not many studies reporting experiments and their replications or aggregating the results of each one of the true experiments to consolidate the findings on chatbot usability.

This project conducts a second replication (with 48 participants) of a baseline experiment (54 participants), which was carried out in July 2019, as well as the first replication (with 30 participants), also conducted in 2019. The ultimate aim is to add together the data from all three experimental studies to ensure the reliability of the results. Meta-analysis techniques is for this purpose.

These experiments evaluate the usability of the SOCIO chatbot against the Creately web application. SOCIO interprets the messages of users of Twitter and Telegram to collaboratively build models or meta-models. Creately is a web application that builds class and other diagrams.

The experiments have a crossover design, in which 50 % of the subjects apply Creately to develop task 1 of the experiment and SOCIO to perform task 2, whereas the remaining 50 % of the subjects apply the treatments in the opposite order. The task is to build class diagrams in groups of three members.

After running the experiment, the results are analysed using boxplots to plot the data related to each of the usability characteristics metrics, that is, effectiveness, efficiency and satisfaction, as well as the quality of the developed models. A linear mixed model is then fitted for each metric. Finally, the effect size produced by the treatment is calculated.

After analysing the original experiment, it was concluded that the sample size was insufficient, and, therefore, none of the metrics generated statistically significant differences. However, the task appeared to generate differences according to all the metrics, except for satisfaction and completion time.

A much larger sample size can be output by aggregating the results of more than one experiment, thereby increasing the statistical evidence to output relevant results. On this ground, we used meta-analysis techniques, which are statistical instruments designed to combine results from families of experiments. Specifically, we used an individual participant data stratified linear regression model.

Finally, the experiment family-level results revealed significant differences between the chatbot SOCIO and the web application Creately for the metrics of quality, satisfaction and efficiency. Specifically, SOCIO was found to have a higher level of efficiency (time to task completion and number of discussion messages exchanged during the task), satisfaction and model precision (percentage of correct items in the solution developed by each team over the items of the ideal diagram), whereas Creately appears to be better in terms of the variable recall (percentage of items of the ideal diagram present in the solution developed by each experimental subject) and perceived success (success rate of each team compared with the ideal solution), both of which are quality metrics.

Thus, the results at the experiment family level were found to be much more accurate than the results of the individual experiment, due to the increased sample size in the aggregation.

Keywords

Chatbot, Usability, Effectiveness, Efficiency, Satisfaction, Crossover Design, Linear Mixed Model, Aggregation

Índice

Listas

Lista de figuras

Lista de tablas

1 Introducción

1.1. Motivación

Los chatbots son elementos de diálogo que se basan en mensajes. Estos se encuentran en pleno crecimiento [9], ya que son utilizados cada vez por un mayor número de usuarios en una gran variedad de ámbitos, tales como ayudando en la búsqueda de información, la planificación de viajes, el control de enfermedades o asistentes personales [4] [14] [24] [23] [12] [3] [10]. Los chatbots han sido diseñados con el objetivo de simplificar las interacciones con los usuarios lo máximo posible.

La usabilidad [7] puede definirse como el nivel en el cual un producto podría ser empleado por clases de usuarios que comparten características particulares, con el objetivo de alcanzar metas específicas relacionadas con características de usabilidad en un contexto particular de uso. Las características habitualmente usadas son eficacia, eficiencia y satisfacción. A través de encuestas, experimentos y tests de usabilidad se está actualmente evaluando la usabilidad de los chatbots [19] [18]. Sin embargo, hay pocos experimentos en la literatura con este objetivo, por lo que el estudio experimental de la usabilidad de chatbots en relación con la satisfacción, eficacia y eficiencia del usuario se considera fuertemente necesario [20] [18] [21].

Es opinión unánime de la comunidad científica que los experimentos singulares poseen escaso valor, con pocas excepciones. La veracidad de los resultados de un experimento base sólo puede establecerse mediante la replicación y contraste de resultados. Una familia de experimentos es un conjunto de replicaciones experimentales donde existe acceso a los datos (brutos o agregados) de cada experimento, se conoce el diseño y protocolo experimental, y contiene al menos tres experimentos con al menos dos tecnologías distintas que ensayan la misma variable respuesta [21]. Las familias de experimentos permiten conseguir un mayor poder estadístico debido al mayor número de sujetos involucrados [13]. De ahí el empleo de técnicas de meta-análisis, que corresponden a instrumentos estadísticos adecuados para combinar resultados de familias de experimentos [1].

SOCIO es un chatbot que interpreta el lenguaje natural en inglés para crear diagramas de clases [16]. Además, SOCIO se encuentra integrado en Telegram y Twitter, permitiendo ser utilizado colaborativamente, en cualquier momento y desde cualquier lugar. Ya se han realizado estudios experimentales para evaluar la usabilidad de SOCIO [20] [18] [11], además de dos evaluaciones a pequeña escala [16] [15]. En cambio, no se encuentran en la literatura réplicas de experimentos verdaderos sobre la usabilidad de SOCIO, ni su agregación.

1.2. Objetivos

La finalidad de este trabajo es realizar la agregación de los datos obtenidos de una familia de experimentos. Los experimentos que conforman esta familia evalúan la usabilidad de SOCIO en relación

con la aplicación web Creately, que permite elaborar, entre otros, diagramas de clases. Se evalúan las siguientes características de usabilidad: eficacia, eficiencia y satisfacción percibida cuando los equipos de tres miembros interactúan con SOCIO, en comparación con Creately. Asimismo, se realiza una valoración de los diagramas de clases diseñados colaborativamente entre los tres integrantes de los equipos involucrados en el experimento. En el Apéndice A se encuentran las tareas y cuestionarios que conforman el experimento.

El experimento original se ha realizado en [20] [18] y la primera réplica en [11]. En este trabajo se realiza la segunda réplica para aplicar las técnicas de meta-análisis con el fin de verificar la consistencia de las conclusiones de los estudios mencionados. Esta réplica es exacta, es decir tanto el diseño del experimento como los modelos estadísticos utilizados durante el análisis de los nuevos datos son los mismos.

Los participantes son estudiantes del Grado conjunto en Ingeniería Informática y Matemáticas perteneciente a la EPS-UAM, así como estudiantes de Ingeniería Informática de la Universidad de las Fuerzas Armadas ESPE (UFA-ESPE) de Ecuador.

1.3. Estructura del trabajo

Este trabajo está compuesto por seis capítulos. Luego de estos, se muestra un apartado con la bibliografía utilizada, un glosario con algunas definiciones importantes y seis apéndices.

En el primer capítulo se detallan las motivaciones y propósitos del trabajo. En el capítulo 2 se estudia el estado de la cuestión de los experimentos realizados con chatbots para evaluar la usabilidad de los mismos, se detallan las propiedades de SOCIO, así como los trabajos relacionados que lo han evaluado, y se presentan las características de la aplicación web Creately, herramienta con la que se compara a SOCIO para la determinación del nivel de usabilidad. En el capítulo 3 se realiza una descripción del experimento llevado a cabo. Se detalla el diseño del experimento y el método de investigación del mismo.

El capítulo cuarto refleja las conclusiones obtenidas al realizar el estudio estadístico sobre la información recopilada durante el experimento y se discuten dichos resultados. El quinto reporta la agregación de los datos correspondientes a: experimento original, primera réplica y segunda réplica (la realizada en este trabajo). En el sexto y último capítulo se exponen las conclusiones de la agregación, realizando una discusión sobre los resultados obtenidos, y se plantean futuros trabajos.

En el Apéndice A se presentan los documentos empleados en el experimento. En el Apéndice B se describen las herramientas utilizadas. En el Apéndice C se explica la evaluación de la calidad y del grado de completitud de los diagramas elaborados en las tareas del estudio experimental.

En el Apéndice D se presentan *boxplots* (en español, diagramas de cajas) para complementar el análisis de SOCIO y Creately. En el Apéndice E se muestra un análisis detallado de los datos particulares de SOCIO, y, en el Apéndice F, el estudio de la normalidad de los datos de la familia de experimentos.

2

Estado del arte

La sección 2.1 de este capítulo analiza una serie de trabajos de investigación que realizan experimentos con chatbots, a partir de un *Systematic Mapping Study* mayor detallado en [19] [18], y a partir del análisis de otros tres trabajos de referencia [13] [21] [22], se evalúa la importancia de las familias de experimentos. En la sección 2.2 se aporta una descripción del chatbot SOCIO. En la sección 2.3 se analizan estudios sobre la evaluación de la usabilidad del chatbot SOCIO. Finalmente, en la sección 2.4 se describe Creately, la herramienta con la que comparamos a SOCIO.

2.1. Evaluación de la usabilidad

En [19] [18] se realizó un *Systematic Mapping Study* (SMS) en el que se identificaron ocho artículos de experimentación con chatbots en diferentes contextos, que se analizan en este trabajo. En el ámbito médico, hay chatbots que ayudaron a controlar enfermedades como la diabetes [4] [23], mientras que otros ofrecen terapia para pacientes con trastornos de estrés pos-traumático [24]. Determinados chatbots facilitan la planificación de viajes [12], el comercio electrónico para la compra de zapatos [9], búsqueda de información [14] o asistentes personales, como por ejemplo Siri o Alexa [3] [10].

Para evaluar la usabilidad en la mayoría de experimentos se compara el chatbot con otro sistema con la misma funcionalidad [4] [9] [12] [14] [23]. Además, las tareas experimentales suelen ser sencillas, como por ejemplo usar Siri para localizar un hotel barato [3] o utilizar el chatbot para reservar una habitación en un hotel y un billete de avión [12].

En tres de los experimentos [9] [12] [14] se aplica un diseño de tipo *within-subject* donde todos los tratamientos a evaluar deben ser aplicados por los sujetos experimentales, y la aplicación de los mismos se da en un orden aleatorio para prevenir los posibles efectos que puedan ser producidos al aplicarlos en un determinado orden. En algunos experimentos, antes de realizar las tareas, los participantes reciben breves tutoriales acerca del chatbot o herramienta que van a aplicar [9] [24].

Finalmente, en todos los experimentos se realizan cuestionarios para obtener información sobre los usuarios y su satisfacción. Estos cuestionarios pueden ser realizados tanto al final del experimento como después de cada una de las tareas y/o antes de empezar el experimento con el fin de conocer la información básica de los sujetos [9] [14].

Mediante la replicación del experimento y posterior agregación de los resultados para formar una familia de experimentos, se consigue un mayor poder estadístico debido al mayor número de sujetos en la familia de experimentos en cuestión [13]. Dicho poder estadístico se traduce en la probabilidad de rechazar la hipótesis nula cuando esta es falsa. Igualmente, las familias de experimentos permiten incrementar también la precisión de los resultados y evaluar el impacto de las variables moderadoras en los mismos [21].

En [21] se realizó un *Systematic Mapping Study* en el que se identificaron las técnicas de agregación empleadas para analizar familias de experimentos, como por ejemplo *Aggregated Data* (AD), *Narrative Synthesis*, *Individual Participant Data stratified* (IPD-S) y *Aggregation of p-values*. Además, según [21] deben mejorarse los informes de análisis de datos para aumentar la fiabilidad y la transparencia de los resultados conjuntos. Se observó que IPD *stratified* y AD parecen ser las técnicas más adecuadas para analizar familias de experimentos en Ingeniería del Software.

En [22] se establecen un conjunto de pautas diseñadas específicamente para el análisis de grupos de réplicas en la Ingeniería del Software. La técnica empleada para agregar grupos de réplicas debe justificarse con el fin de, como se ha mencionado, incrementar la fiabilidad y la transparencia de los resultados conjuntos. Asimismo, según [22] se debe aprovechar toda la información contenida en los datos sin procesar de cada experimento de la familia, y se fomenta el uso de IPD *stratified* y AD para analizar grupos de réplicas.

En [13] se lleva a cabo un estudio de una familia de experimentos mediante dos tipos de réplicas: unas similares al experimento inicial (*strict replications*) y otras en las que aumentaron la complejidad del problema (*object replications*). El objetivo era evaluar el impacto que pudiese tener la complejidad del problema en la calidad del software en la ingeniería dirigida por modelos. Gracias al gran tamaño muestral obtenido tras agregar las réplicas, se obtuvieron resultados que en el experimento inicial no habían surgido debido a un pequeño tamaño muestral.

En IPD *stratified* [22] se deben analizar todos los datos de los experimentos que forman la familia conjuntamente, reconociendo el experimento del que proceden. Se pueden ajustar modelos de efectos fijos o modelos de efectos aleatorios. Para los modelos de efectos fijos se suelen emplear modelos de regresión lineal de dos factores: experimento y tratamiento. Para los modelos de efectos aleatorios se suele ajustar un modelo lineal mixto también de dos factores: experimento y tratamiento.

Vista la importancia de las familias de experimentos y los escasos estudios sobre ellas encontrados en la literatura, en este trabajo hemos realizado una réplica y la posterior agregación con dos experimentos anteriores [20] [18] [11]. La técnica que hemos utilizado para realizar la agregación es IPD *stratified*, ya que incrementa la interpretabilidad de los resultados conjuntos y ofrece gran estabilidad estadística.

2.2. Breve descripción de SOCIO

El chatbot SOCIO (`https://saraperezsoler.github.io/ModellingBot/`) [16] asiste en el diseño de diagramas de clases interpretando mensajes en lenguaje natural en inglés, y se encuentra integrado en Telegram y Twitter (redes sociales a las que los participantes están acostumbrados), permitiendo ser utilizado colaborativamente.

Los usuarios pueden comunicarse entre ellos de la manera habitual en las redes sociales. Del mismo modo, los mensajes pueden ir dirigidos a SOCIO, quien los interpretará para construir el diagrama. En la red social de Twitter, los usuarios tienen que seguir a SOCIO y, cada vez que quieran dirigirle un mensaje, deben citar al bot mediante su nombre de usuario (@modellingBot).

En el experimento hemos utilizado esta herramienta mediante Telegram. En Telegram podemos comunicarnos con SOCIO bien mediante un chat o bien mediante un grupo de chat del que el chatbot

forme parte, nuevamente su alias es @modellingBot. En nuestro caso, como el experimento se ha llevado a cabo por equipos formados por tres participantes cada uno, hemos optado por crear grupos de Telegram para su interacción con el bot.

SOCIO admite comandos de gestión, como crear un proyecto nuevo o acceder a las contribuciones de los usuarios, entre otros. Asimismo, puede interpretar dos tipos de mensajes para la actualización de los diagramas: mensajes descriptivos o comandos. Por un lado, los mensajes descriptivos son oraciones en lenguaje natural, como por ejemplo "*the house contains rooms*". Por otro lado, los comandos son órdenes, por ejemplo para añadir una clase o cambiar el tipo de un atributo. En el Apéndice B se muestran algunos ejemplos de estos comandos y mensajes descriptivos.

2.3. Trabajos relacionados sobre SOCIO

En [16] y [15] se presentan dos evaluaciones del chatbot SOCIO a pequeña escala. Detalles de las mismas se pueden apreciar en la Tabla 2.1.

	[16]	**[15]**
Evaluación	Idoneidad de SOCIO	Mecanismo de consenso
Participantes	10	8
Grupos de Telegram	4 (de 2 y 3 personas)	1 (con todos los participantes)
Tareas	Diagrama de clases para el comercio electrónico (15 minutos máximo)	Tras un pequeño tutorial, se debe elegir una de 3 alternativas considerando 2 proyectos: en primer lugar, teniendo en cuenta la ausencia del mecanismo de consenso, y posteriormente su presencia
Cuestionarios	Tras realizar las tareas, se centra en la satisfacción, integración en redes sociales y uso del lenguaje natural	Tras las tareas, se focaliza en el mecanismo de consenso
Resultados	La satisfacción presenta resultados positivos, aunque se refleja una necesidad en la mejora de la interpretación del lenguaje natural	El mecanismo de consenso fue considerado positivo por la mayoría

Tabla 2.1: Evaluaciones de SOCIO.

En este trabajo realizaremos agregaciones de datos y su análisis, partiendo del experimento base [20] [18], la primera de sus réplicas [11] y esta segunda réplica, para comprobar la consistencia de los resultados y la obtención de evidencias empíricas consolidadas a fin de mejorar la usabilidad del chatbot SOCIO.

El diseño del experimento y los modelos estadísticos utilizados durante el análisis de los datos son los mismos en cada uno de los experimentos singulares que conforman esta familia, tal y como exigen las réplicas exactas de un experimento base.

Por una parte, con respecto al experimento original [20] [18], SOCIO tuvo una puntuación ventajosa (con respecto a Creately) en cuanto a la satisfaccion, eficacia y eficiencia. En cuanto a la calidad de los modelos creados, SOCIO obtuvo mejores puntuaciones de precisión, aunque Creately obtuvo mejores puntuaciones de recall y aciertos. En resumen, la usabilidad de SOCIO tuvo un efecto positivo en la mayoría de aspectos.

Por otra parte, en la primera réplica [11] se obtuvo una mayor satisfacción con la aplicación del chatbot SOCIO en relación con la de Creately, pero los resultados para la eficacia, la eficiencia y la calidad fueron similares para ambos tratamientos (SOCIO y Creately), por lo que no se obtuvieron resultados concluyentes.

2.4. Creately

Creately (`https://creately.com/app/`) es una herramienta colaborativa para crear diagramas de múltiples tipos, entre ellos están los diagramas de clases. A través de la dirección de correo electrónico, otros usuarios pueden ser invitados a participar en la elaboración del diagrama. En Creately se trabaja de forma online (aunque también permite trabajar offline y volver a sincronizar cuando haya conexión). En el Apéndice B se muestra una explicación para el desarrollo de diagramas de clases a través de la herramienta Creately.

Existen muchas herramientas colaborativas para elaborar diagramas, Creately es una de las más utilizadas [20] [18]. Además, al no haber estudios previos acerca de la usabilidad de Creately, y al tener una funcionalidad similar que la de SOCIO, se elige como herramienta de comparación en esta familia de experimentos para nuestro estudio de usabilidad.

Sin embargo, Creately carece de un chat mediante el que los integrantes de los equipos que realizan el diagrama puedan comunicase, por lo que en nuestro experimento se utiliza un chat externo (Telegram) en su lugar.

3

Experimento

Este capítulo presenta la configuración de la réplica llevada a cabo para la evaluación de la usabilidad del chatbot SOCIO en relación con la aplicación web Creately. En las siguientes secciones se explican el diseño del experimento (sección 3.1), sus objetivos e hipótesis (sección 3.2), los factores y variables respuesta del mismo (sección 3.3), los sujetos experimentales (sección 3.4), las herramientas y tareas realizadas (sección 3.5), el desarrollo del experimento (sección 3.6) y, para finalizar, las amenazas a la validez que se presentan (sección 3.7).

Nótese que este experimento es una réplica de un experimento base [18] [20] cuyo diseño es el mismo en cada uno de los experimentos que conforman la familia, tal y como exigen las réplicas exactas de un experimento original.

3.1. Diseño experimental

Se trata de un experimento con un diseño de tipo **crossover within-sujects**. En esta clase de diseños, cada sujeto experimental aplica todos los tratamientos aunque en distinto orden [25].

En nuestro experimento los dos **tratamientos** aplicados son SOCIO y Creately, ambas herramientas, chatbot y aplicación web respectivamente, permiten la creación de diagramas de clases.

Los sujetos experimentales fueron divididos, aleatoriamente, en dos grupos, A y B. Dentro de cada uno de estos grupos, se forman equipos de tres individuos elegidos al azar. Los equipos pertenecientes al grupo A aplican Creately primero y SOCIO después, y los equipos pertenecientes al grupo B aplican los tratamientos en el orden inverso (SOCIO primero y Creately después). Esto es lo que se denominan **secuencias**, es decir, los órdenes de aplicación de cada tratamiento.

El momento de aplicación de cada tratamiento son los **periodos**. Los participantes aplican cada tratamiento una sola vez, por lo que hay dos periodos. En el primer periodo se aplica el primer tratamiento (según el grupo al que pertenezca el equipo en cuestión) a la primera tarea, mientras que en el segundo periodo (nuevamente, según el grupo experimental) el segundo tratamiento es aplicado a la segunda tarea.

En la Tabla 3.1 se muestra el diseño experimental seguido.

	Periodo 1	**Periodo 2**	
	Tarea 1	**Tarea 2**	
Grupo A	Creately	SOCIO	**Secuencias**
Grupo B	SOCIO	Creately	

Tabla 3.1: Diseño experimental.

3.2. Hipótesis y objetivo de la investigación

El propósito de este trabajo consiste en establecer el nivel de usabilidad del chatbot SOCIO comparándolo con la aplicación web Creately, según las métricas correspondientes a la eficacia, a la eficiencia, a la satisfacción de los usuarios y a la calidad del diagrama de clases desarrollado durante el experimento, empleando las mencionadas herramientas.

De esta manera, la pregunta de investigación es:

RQ: ¿Ejerce la aplicación del chatbot SOCIO una influencia significativa en la eficacia, en la eficiencia y en la satisfacción de los equipos, además de en la calidad de los diagramas, con respecto a Creately?

Las hipótesis tanto para SOCIO como Creately son las siguientes:

H.1.0: No hay relación entre elaborar el diagrama con SOCIO o con Creately con respecto a la eficacia.

H.2.0: No hay relación entre elaborar el diagrama con SOCIO o con Creately con respecto a la eficiencia.

H.3.0: No hay relación entre elaborar el diagrama con SOCIO o con Creately con respecto a la satisfacción.

H.4.0: No hay relación entre elaborar el diagrama con SOCIO o con Creately con respecto a la calidad.

No obstante, el empleo de SOCIO aporta mayor información que Creately, por lo que se consideran las siguientes hipótesis de acuerdo a los datos proporcionados por el chatbot:

H.S.1.0: No hay relación entre las tareas 1 y 2 con respecto a la cantidad de mensajes enviados a SOCIO.

H.S.2.0: No hay relación entre las tareas 1 y 2 con respecto a la cantidad de mensajes erróneos cometidos por los equipos y enviados a SOCIO.

H.S.3.0: No hay relación entre las tareas 1 y 2 con respecto a la cantidad de mensajes válidos interpretados erróneamente por SOCIO.

H.S.4.0: No hay relación entre las tareas 1 y 2 con respecto a la cantidad de mensajes de error dirigidos a SOCIO.

H.S.5.0: No hay relación entre las tareas 1 y 2 con respecto a la cantidad de mensajes útiles dirigidos a SOCIO.

H.S.6.0: No hay relación entre las tareas 1 y 2 con respecto a la cantidad de mensajes descriptivos enviados a SOCIO.

H.S.7.0: No hay relación entre las tareas 1 y 2 con respecto a la cantidad de comandos enviados a SOCIO.

H.S.8.0: No hay relación entre las tareas 1 y 2 con respecto a la cantidad de acciones desencadenadas por SOCIO.

3.3. Variables respuesta y factores

Los **factores** [25] de esta clase de experimentos con un diseño *crossover* son: el **tratamiento** (herramienta aplicada por los equipos para realizar las tareas), la **secuencia** (orden en el que se aplican los tratamientos) y el **periodo** (momento de aplicación de los tratamientos).

En el caso de nuestro experimento, cada uno de los factores presenta dos categorías. SOCIO y Creately son los dos tratamientos aplicados, chatbot y aplicación web respectivamente. Las dos secuencias son Creately-SOCIO (para los equipos del grupo A) y SOCIO-Creately (para los equipos del grupo B), y, para finalizar, hay dos periodos, el periodo 1 (cuando se aplica el primer tratamiento) y el periodo 2 (cuando se aplica el segundo de los tratamientos). La tarea 1 se realiza en el primer periodo y la tarea 2 en el segundo periodo, de modo que los efectos producidos por el factor periodo pueden ser confundidos con los efectos producidos por la tarea.

Con respecto a las **variables respuesta**, en [8] se definen la **eficacia**, la **eficiencia** y la **satisfacción** como métricas estándar para evaluar la usabilidad. A dichas métricas se añade en este estudio la **calidad**, con el fin de medir la calidad de los diagramas obtenidos.

Las métricas que presentan un asterisco de las detalladas a continuación son las que corresponden a ambas herramientas (SOCIO y Creately), mientras que las que no tengan asterisco se corresponden con los datos particulares de SOCIO.

La **eficacia** se mide a través del nivel de **completitud** de las tareas desarrolladas por los equipos $(*)$. En el Apéndice C se especifica el cálculo del grado de completitud de los diagramas.

Las **métricas** que miden la **eficiencia** para cada uno de los equipos a lo largo de cada tarea son:

- **Rapidez.**
  - **Tiempo** en terminar una tarea, se representa en minutos y el máximo son 30. $(*)$.
- **Fluidez.**
  - **Nº de mensajes intercambiados de discusión** por los integrantes de un equipo para comunicarse entre ellos $(*)$.
  - **Nº de mensajes dirigidos al chatbot.**
  - **Nº de mensajes erróneos por fallos cometidos por los equipos dirigidos al chatbot SOCIO.**
  - **Nº de mensajes correctos enviados al chatbot aunque erróneamente interpretados por SOCIO.**
  - **Nº de mensajes erróneos enviados a SOCIO.** Incluyen tanto los mensajes erróneos por fallos de los equipos como los mensajes correctos interpretados erróneamente por SOCIO.
- **Interactividad.**
  - **Nº de mensajes útiles enviados a SOCIO.** Son los mensajes que han contribuido al diagrama resultante. Engloban a los mensajes descriptivos y a los comandos.

- **Nº de mensajes descriptivos** dirigidos al bot.
- **Nº de comandos** dirigidos al bot.
- **Nº de acciones desencadenadas por SOCIO** en relación a los mensajes que han contribuido al diagrama.

Mediante las respuestas de los participantes al cuestionario SUS (*System Usability Scale*) [2], se mide la **satisfacción** de los mismos. El valor de cada una de las respuestas se corresponde a un número ordinal de la escala de Likert ("totalmente en desacuerdo" para el valor 1 o "totalmente de acuerdo" para el valor 5). El cálculo de la satisfacción de cada equipo se realiza mediante la mediana de las respuestas de sus integrantes. Para obtener la puntuación de satisfacción final se realiza el promedio de las puntuaciones de satisfacción de los equipos $(*)$.

La **calidad** de las soluciones a las tareas desarrolladas por parte de los sujetos experimentales se mide con respecto a la solución ideal, mostrada en el Apéndice A. Las métricas empleadas para medir la calidad son las siguientes:

- Precisión = $\frac{TP}{TP+FP}(*)$.
- Recall = $\frac{TP}{TP+FN}(*)$.
- Accuracy = $\frac{TN+TP}{TP+TN+FP+FN}(*)$.
- Aciertos = $\frac{TP}{N^{\circ}\ elementos\ diagrama\ ideal}(*)$.
- Error = $\frac{FP+FN}{TP+TN+FP+FN}(*)$.

Donde TP (*true positive*), FP (*false positive*), TN (*true negative*) y FN (*false negative*) se obtienen comparando cada uno de los diagramas realizados por los equipos con el diagrama ideal. En el Apéndice C se especifica más detalladamente el cálculo de cada una de estas métricas.

3.4. Sujetos del experimento

Para la realización del experimento se ha contado con la participación voluntaria de un total de 48 estudiantes tanto del Doble Grado en Ingeniería Informática y Matemáticas, como del grado de Ingeniería Informática, formados en la EPS-UAM y en la UFA-ESPE de Ecuador, respectivamente. Todos los participantes han tenido las materias de Análisis y Diseño de Software, así como Proyecto de Análisis y Diseño de Software o similares, de modo que todos ellos poseen los dominios fundamentales para realizar un diagrama de clases.

Los participantes son agrupados aleatoriamente en equipos de tres integrantes, y cada uno de estos equipos se asigna al azar a un grupo experimental (los grupos experimentales son los mostrados en la Tabla 3.1). Antes de aplicar cada uno de los tratamientos reciben un breve tutorial sobre los mismos. A continuación, se detallan las características de los participantes obtenidas a partir de las respuestas al cuestionario de familiaridad.

- Se cuenta con un total de 48 participantes, de los cuales 31 son hombres (65 %) y 17 mujeres (35 %). Todos ellos de entre 20 y 26 años de edad, con una media de 22,31 años y una desviación típica de 1,2 años.

- El 79 % de los participantes consideran que tienen un conocimiento medio-alto sobre diagramas

de clases. Con respecto al nivel de inglés, solo el 10 % consideran tener un nivel bajo.

• Los sujetos están acostumbrados al uso de las redes sociales, en concreto el 81 % indican hacer un uso medio-alto de ellas. Los participantes utilizan principalmente WhatsApp (97 %), mientras que el uso de Telegram, Twitter, Facebook e Instagram está más equilibrado entre los sujetos.

• La gran mayoría de los participantes, concretamente un 87 %, han utilizado Telegram alguna vez, red social utilizada en este estudio experimental. El 73 % de los usuarios realizará el experimento a través de la aplicación móvil de Telegram, mientras que el 27 % restante utilizarán la aplicación web.

• El 35 % de los participantes señalan no haber utilizado nunca un chatbot, el 78 % indican que usan los chatbots con un grado nulo o bajo, y el 73 % consideran que su grado de conocimiento sobre chatbots es bajo o muy bajo.

3.5. Tareas y herramientas

El experimento ha sido llevado a cabo a través de SOCIO y Creately, chatbot y aplicación web respectivamente. Ambas son las herramientas utilizadas en el experimento.

SOCIO es un chatbot que permite elaborar diagramas de clases, interpretando mensajes escritos en lenguaje natural en inglés. Además, SOCIO se encuentra integrado en Telegram y Twitter, permitiendo ser utilizado de forma colaborativa. En este estudio, los participantes han sido agrupados aleatoriamente en equipos de tres integrantes, los mencionados equipos han interactuado con SOCIO mediante un grupo de Telegram, el cual se constituye por cada uno de los miembros del equipo y SOCIO, cuyo alias es @modellingBot.

La aplicación web Creately (`https://creately.com/app/`) igualmente permite la creación de diagramas de clases de forma colaborativa. Además, también posibilita la realización de otros múltiples tipos de diagramas.

Previamente a la aplicación de los tratamientos, cada participante debe rellenar unos documentos iniciales (informe de consentimiento y cuestionario de familiaridad) como los del Apéndice A. Después, cada uno de los equipos aplican cada herramienta en una de las dos tareas del experimento, mostradas también en el Apéndice A.

En la **tarea 1**, cada equipo debe realizar un diagrama de clases (únicamente con las clases y atributos, no los métodos) de una aplicación para una tienda, para tratar tanto sus clientes como sus productos. Tal y como se muestra en la Tabla 3.1 los equipos del grupo A deben llevar a cabo esta tarea con la herramienta Creately, mientras que los del grupo B con la herramienta SOCIO.

En la **tarea 2**, cada equipo debe realizar un diagrama de clases (únicamente con las clases y atributos, no los métodos) de una aplicación para un colegio, con el fin de tratar sus estudiantes, profesores y asignaturas. Como se muestra en la Tabla 3.1 los sujetos experimentales del grupo A deben desempeñar esta tarea con la herramienta SOCIO, mientras que los del grupo B la realizan con la herramienta Creately.

Hay un tiempo máximo de 30 minutos para realizar las tareas. Durante la realización no está permitido hablar en voz alta, los integrantes de cada equipo deben comunicarse entre ellos mediante su

grupo de Telegram. Tras cada tarea, cada uno de los participantes debe rellenar un cuestionario SUS (mostrado en el Apéndice A) acerca de la herramienta que acaban de aplicar.

3.6. Operación

El experimento fue realizado en noviembre de 2019, de modo presencial en los laboratorios 6A, 2 y 12, y de modo remoto (con estudiantes de la UFA-ESPE en Ecuador) en el seminario B-351 de la EPS-UAM.

Dicho experimento ha sido dividido en 10 sesiones. Según la disponibilidad de los participantes se les ha asignado a una sesión u otra. La Tabla 3.2 representa el reparto de los equipos según el grupo y la sesión que le corresponde.

Fecha	Sesión	Equipos	Grupo
11/11/2019	1	1	A
13/11/2019	3	5, 6	
13/11/2019	4	7	
15/11/2019	6	9	
18/11/2019	8	11	
22/11/2019	9	15, 16	
12/11/2019	2	2, 3, 4	B
14/11/2019	5	8	
15/11/2019	7	10	
22/11/2019	10	12, 13, 14	

Tabla 3.2: Sesiones, sujetos experimentales y grupos del estudio.

En todas las sesiones, se ha distribuido a los participantes por los laboratorios para que estuviesen separados con el fin de impedir copias entre ellos y que se comunicasen en voz alta. En Ecuador se disponía de la colaboración del profesor responsable de la asignatura equivalente a Análisis y Diseño de Software para lograr este fin.

Al empezar el experimento, cada uno de los participantes debía firmar el acuerdo de consentimiento y rellenar un cuestionario para recoger información básica acerca de los sujetos.

Después, según el grupo correspondiente a cada equipo, los participantes recibieron un tutorial sobre la herramienta que aplicarían en la tarea 1. Esta primera tarea se realizaba con Creately en el caso del grupo A, y con SOCIO en el grupo B. Tras su realización, los participantes rellenaban individualmente el cuestionario de satisfacción (SUS) asociado a la herramienta que acababan de aplicar.

Finalmente, del mismo modo, tras impartir el tutorial pertinente a la segunda herramienta según el grupo experimental, se llevaba a cabo la segunda de las tareas, el grupo A la realizaba con SOCIO mientras que el grupo B realizaba la misma tarea utilizando Creately. Tras su realización, cada participante debía completar el cuestionario de satisfacción correspondiente a la última herramienta aplicada.

Todos los documentos mencionados para el experimento se muestran en el Apéndice A. En la Figura 3.1 se detalla este proceso.

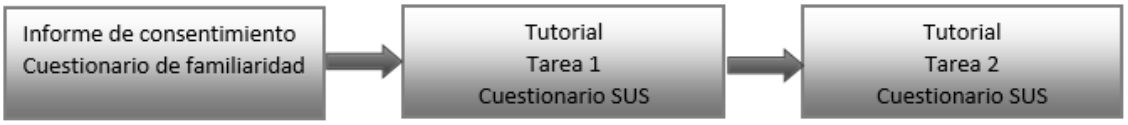


Figura 3.1: Proceso para la realización del experimento.

3.7. Amenazas a la validez

En los experimentos con un diseño *crossover* como el nuestro las amenazas a la **validez interna** pueden ser causadas por el número y distribución de los periodos, así como la selección de las secuencias [25]. Esta validez hace referencia al nivel de confianza de las conclusiones obtenidas, es decir, que sean válidas e interpretables.

Las amenazas causadas por el periodo son:

• Hay amenaza de aprendizaje por la práctica. En el experimento, los participantes deben realizar dos diagramas de clases y, a pesar de que ya saben cómo realizarlos, en la elaboración del primero de los diagramas refrescarán los conocimientos y recuperarán la práctica, por lo que la aplicación del segundo tratamiento puede producir mejores resultados. La forma de mitigar esta amenaza consiste en comparar los resultados obtenidos en ambos periodos y estudiar las mejoras observadas.

• No hay amenaza de copia entre periodos, ya que ambos se llevan a cabo de forma continuada en la misma sesión y las tareas son diferentes. Sin embargo, como hay múltiples sesiones, existe una amenaza de copia entre los participantes si éstos comentan las tareas que han realizado con otros sujetos de sesiones posteriores. Para mitigar esta amenaza no se informó a los participantes de que las tareas serían las mismas para todas las sesiones, y se les pidió que no comentasen nada tras finalizar la sesión a la que habían sido asignados, hasta que finalizasen todas las sesiones del experimento.

• Existe una amenaza de cansancio o aburrimiento ya que las sesiones duran una hora y media. Además, los participantes pueden presentar una falta de motivación ya que se trata de un experimento voluntario que no les supone ninguna repercusión en las calificaciones.

En cuanto a las amenazas causadas por la secuencia, no se considera que pueda haber una secuencia óptima en la aplicación de los tratamientos mediante la cual se obtengan mejores resultados.

El *carryover* es otra de las amenazas que presentan los experimentos con un diseño *crossover*. El *carryover* se produce al aplicar un tratamiento previamente a que haya desaparecido el efecto del tratamiento anterior. De este modo, en caso de que los primeros tratamientos mejoren la efectividad de los posteriores, los tratamientos aplicados en primer lugar pueden parecer menos efectivos en comparación, mientras que si los primeros tratamientos disminuyen la efectividad pueden parecer más efectivos que los aplicados posteriormente.

Además, en este tipo de experimentos en los que hay igual número de tratamientos que de periodos, puede haber confusión entre el *carryover*, la interacción tratamiento-periodo y los efectos producidos por la secuencia, por lo que es imposible distinguir cuál de los tres está ocurriendo realmente [25].

La **validez externa** se refiere a la manera en la que las conclusiones del experimento pueden ser generalizadas a otros ámbitos. Los usuarios que han participado en este estudio experimental son estudiantes de informática con conocimientos suficientes para realizar diagramas de clases, por lo tanto, los resultados permanecen en el ámbito académico y no son generalizables a otros campos.

Las amenazas a nivel de la familia de experimentos están relacionadas con la **validez de las conclusiones**. Son las amenazas que aparecen al replicar el experimento y agregar los resultados. En este caso, hay una amenaza debido a un bajo poder estadístico, ya que como el experimento se realiza en equipos de tres participantes, contamos con un menor número de sujetos experimentales (44 equipos) que otras familias de experimentos, en las que suele haber cerca de 100 sujetos [13].

4

Enfoque de análisis

El presente capítulo detalla el análisis estadístico de la información recopilada a lo largo de la ejecución del experimento definido anteriormente. La sección 4.1, reporta el estudio estadístico de la información asociada tanto a SOCIO como a Creately, para evaluar las métricas de las características de usabilidad (eficacia, eficiencia y satisfacción) y la calidad de los diagramas desarrollados por los sujetos experimentales con las mencionadas herramientas. En la sección 4.2, se detalla la discusión de los resultados de este análisis.

4.1. Análisis

Ha sido realizado un análisis gráfico utilizando *boxplots* para los datos recopilados durante el experimento asociados a ambas herramientas, SOCIO y Creately. Dichos *boxplots* representan la información, agrupada por tratamiento, asociada a cada una de las métricas. Adicionalmente, en el Apéndice D se muestran los diagramas pertinentes a la información recopilada y agrupada por tratamiento-periodo/tarea, y por tratamiento-secuencia.

A raíz de este estudio descriptivo, se procesa la información recogida mediante el modelo lineal mixto (en inglés *Linear Mixed Model*, LMM) para cada una de las métricas de este análisis, según [25]. Dicho modelo es el mejor método como extensión del modelo lineal generalizado para analizar modelos con coeficientes aleatorios (los sujetos experimentales) y datos dependientes por medidas repetidas (se trata de un diseño *crossover*, por lo que en cada aplicación de los tratamientos por parte de los sujetos se toman las mismas medidas). Todos los modelos presentados conllevan iguales factores:

• Secuencia: es el orden de aplicación de un tratamiento u otro, en nuestro caso hay dos secuencias, Creately-SOCIO o SOCIO-Creately. En diseños de tipo *crossover* como es el caso, los efectos de la secuencia pueden confundirse con los de la interacción tratamiento-periodo y los del *carryover* [25].

• Tratamiento: es el instrumento utilizado por parte de los sujetos experimentales para la realización de las tareas. En nuestro caso hay dos tratamientos, Creately o SOCIO.

• Periodo: se refiere a la aplicación de los tratamientos a las tareas por parte de los equipos, en nuestro caso hay dos periodos. En el primer periodo, cuando se aplica el primer tratamiento, se desarrolla la primera tarea, mientras que en el segundo periodo, cuando se aplica el segundo de los tratamientos, se desarrolla la segunda tarea. Por ello, los efectos del periodo y los de la tarea pueden ser confundidos.

Otro de los factores adicionales a tener en cuenta a lo largo del análisis desarrollado deriva en el tamaño del efecto producido por el tratamiento, que mide la diferencia entre ambos tratamientos. El método empleado para medir la mencionada magnitud es a través del cálculo de la d de Cohen (la cual llamaremos d desde este momento), y su correspondiente error estándar (SE) [6].

4.1.1. Eficacia

Como métrica para valorar la eficacia de las tareas se usa el nivel de completitud. En la Figura 4.1 se representa el *boxplot* de dicha métrica, agrupado por tratamiento. Como podemos ver en la Figura 4.1, los resultados para el grado de completitud parecen similares para ambos tratamientos, SOCIO y Creately.

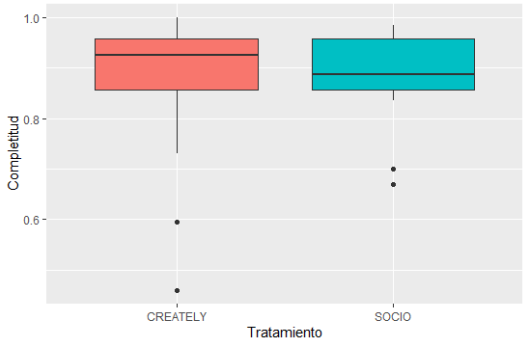


Figura 4.1: *Boxplot* de las valoraciones de completitud.

Los resultados del LMM ajustado en el análisis son presentados en la Tabla 4.1. Como podemos apreciar en la Tabla 4.1, el periodo es un factor estadísticamente significativo. En la primera tarea se observa que la puntuación de completitud es 0.074 veces mayor que en la segunda tarea. En cuanto al tratamiento, Creately obtiene de media 0.014 puntos más que SOCIO, lo cual no es significativo en comparación con los 0.92 puntos de media. Por último, d=0.1, SE(d)= 0.25, lo que indica que el tamaño del efecto es pequeño, según [1].

	Estimate	Std.Error	p-value
(Intercept)	0.92	0.046	0
Seq	-0.035	0.057	0.56
Treatment	0.014	0.03	0.65
Period	-0.074	0.03	**0.027**

Tabla 4.1: LMM de la completitud.

En definitiva, **ambos tratamientos obtienen puntuaciones similares en cuanto a la completitud.**

4.1.2. Eficiencia

La eficiencia se mide en términos de la velocidad y la fluidez. La velocidad se corresponde con el tiempo en llevar a cabo las tareas, mientras que la fluidez hace referencia a la cantidad de mensajes intercambiados de discusión entre los integrantes de cada uno de los equipos a lo largo de las tareas.

Tiempo

En la Figura 4.2 se representa el *boxplot* referente al tiempo en que los sujetos experimentales realizan las tareas, donde los datos se agrupan por tratamiento. Como podemos ver, parece que ambos tratamientos requieren tiempos similares.

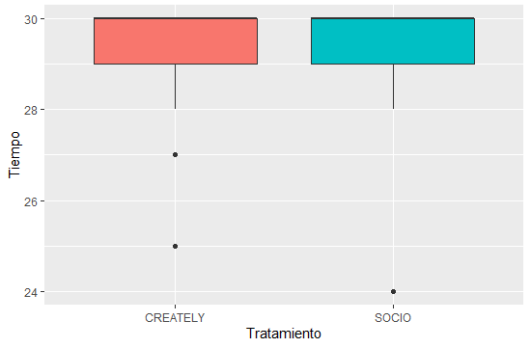


Figura 4.2: *Boxplot* del tiempo utilizado para la realización de la tarea.

Los resultados del LMM ajustado en el análisis son presentados en la Tabla 4.2. Como podemos apreciar en la Tabla 4.2, no hay factores estadísticamente significativos. Se obtienen tiempos muy similares tanto para SOCIO como para Creately, con una media de 29.13 minutos en la realización de las tareas. Por último, d=0, SE(d)=0.34, de modo que no hay ningún efecto, según [1].

	Estimate	Std.Error	p-value
(Intercept)	29.13	0.54	0
Seq	-0.13	0.56	0.83
Treatment	0	0.52	1
Period	0.25	0.52	0.64

Tabla 4.2: LMM del tiempo utilizado para realizar la tarea.

En definitiva, **ambos tratamientos requieren un tiempo similar en realizar las tareas.**

Número de mensajes de discusión

El *boxplot* referente a los mensajes intercambiados de discusión se representa en la Figura 4.3, donde los datos se agrupan por tratamiento. Como podemos ver, hay un número de mensajes similar para ambos tratamientos, aunque las puntuaciones del número de mensajes para SOCIO se encuentran más dispersas.

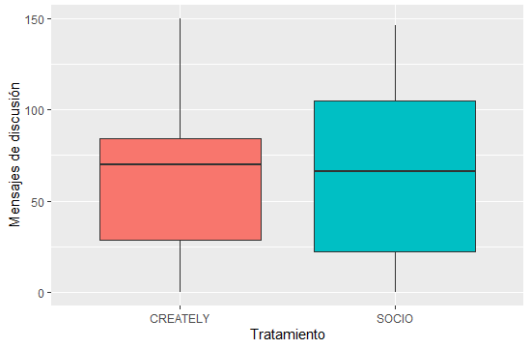


Figura 4.3: *Boxplot* del nº de mensajes de discusión.

Los resultados del LMM ajustado en el análisis son presentados en la Tabla 4.3. Como podemos apreciar en la Tabla 4.3, el factor estadísticamente significativo es el periodo, lo cual afecta en el número de mensajes intercambiados de discusión. En la segunda tarea observamos que la puntuación para

el número de mensajes intercambiados de discusión es 16.19 veces menor con respecto a la primera tarea. Por último, d=0.061, SE(d)=0.15, de modo que el tamaño del efecto es pequeño, según [1].

	Estimate	Std.Error	p-value
(Intercept)	78.63	16.28	0
Seq	-15.06	22.25	0.51
Treatment	2.81	5.93	0.64
Period	-16.19	5.93	**0.016**

Tabla 4.3: LMM del nº de mensajes intercambiados de discusión entre los integrantes de los equipos.

En definitiva, **para ambos tratamientos se obtiene un número de mensajes similar, requiriendo dichos tratamientos un esfuerzo muy parecido.**

4.1.3. Satisfacción

En la Figura 4.4 se representa el *boxplot* referente a las valoraciones de satisfacción de los sujetos experimentales, donde los datos se agrupan por tratamiento. Como podemos ver, para los dos tratamientos se tienen puntuaciones similares.

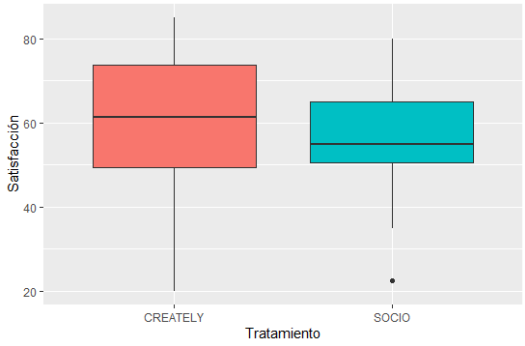


Figura 4.4: *Boxplot* de las valoraciones de satisfacción.

Los resultados del LMM ajustado en el análisis son presentados en la Tabla 4.4. Como podemos apreciar en la Tabla 4.4, no hay factores estadísticamente significativos. Creately obtiene una puntuación de satisfacción 4.53 mayor que SOCIO, lo cual no es significativo teniendo en cuenta que de media hay 57.81 puntos de satisfacción. Por último, d=-0.27, SE(d)= 0.31, de modo que el tamaño del efecto es pequeño, según [1].

	Estimate	Std.Error	p-value
(Intercept)	57.81	5.53	0
Seq	11.72	5.98	0.07
Treatment	-4.53	5.05	0.385
Period	-7.03	5.05	0.186

Tabla 4.4: LMM de la satisfacción.

En definitiva, **ambos tratamientos parecen satisfacer por igual a los equipos.**

4.1.4. Calidad

Para cada uno de los diagramas desarrollados por parte de los sujetos experimentales, en función de varios aspectos (métricas): accuracy, precisión, recall, aciertos y error, se ha realizado el análisis de la calidad. En el grupo de investigación se ha decidido no traducir accuracy y recall por su similitud con otros términos como “recuperación” en relación con otros ámbitos.

Accuracy

En la Figura 4.5 se representa el *boxplot* referente a las valoraciones de accuracy asociados a los diagramas realizados por los sujetos experimentales mediante los tratamientos SOCIO y Creately. Como podemos contemplar, las valoraciones para SOCIO se encuentran por encima que las valoraciones para Creately.

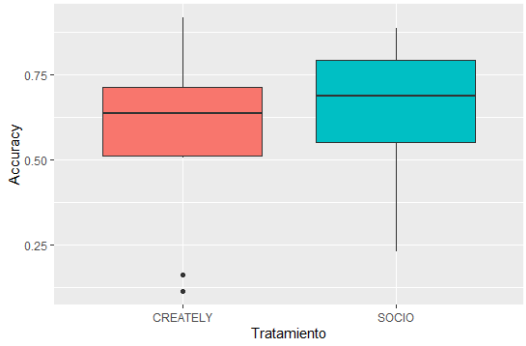


Figura 4.5: *Boxplot* de las valoraciones de accuracy.

Los resultados del LMM ajustado en el análisis son presentados en la Tabla 4.5. Como podemos apreciar en la Tabla 4.5, el factor estadísticamente significativo es el periodo. Las valoraciones de accuracy en la segunda tarea se encuentran por debajo que las de la primera tarea, con una diferencia de 0.19 puntos. Por último, d=0.11, SE(d)=0.31, de modo que se tiene un pequeño tamaño del efecto, según [1].

	Estimate	Std.Error	p-value
(Intercept)	0.72	0.07	0
Seq	-0.03	0.08	0.74
Treatment	0.02	0.05	0.62
Period	-0.19	0.05	**0**

Tabla 4.5: LMM de la variable accuracy.

En definitiva, **ambos tratamientos parecen obtener puntuaciones similares de accuracy.**

Precisión

En la Figura 4.6 se representa el *boxplot* referente a las valoraciones de precisión de los diagramas llevados a cabo utilizando SOCIO y Creately por parte de los sujetos experimentales. Como podemos ver, se obtienen valoraciones de precisión parecidas con los dos tratamientos.

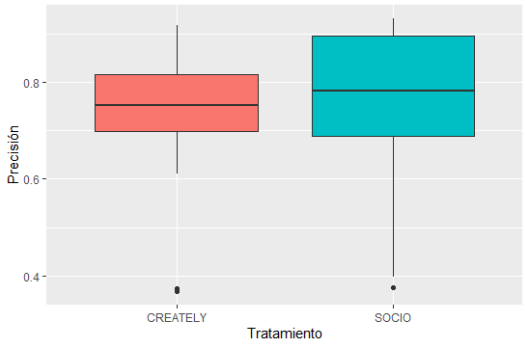


Figura 4.6: *Boxplot* de las valoraciones de precisión.

Los resultados del LMM ajustado en el análisis son presentados en la Tabla 4.6. Como podemos apreciar en la Tabla 4.6, el factor estadísticamente significativo es el periodo. Para la primera tarea las valoraciones de precisión son mayores que para la segunda, en concreto, con una diferencia de 0.14 puntos. Por último, d=0.08, SE(d)=0.34, por lo que se tiene un pequeño tamaño del efecto, según [1].

	Estimate	Std.Error	p-value
(Intercept)	0.81	0.057	0
Seq	-0.012	0.065	0.85
Treatment	0.014	0.05	0.77
Period	-0.14	0.05	**0**

Tabla 4.6: LMM de la variable precisión.

En definitiva, **ambos tratamientos parecen obtener puntuaciones de precisión muy similares.**

Recall

En la Figura 4.7 se representa el *boxplot* referente a las valoraciones de recall de los diagramas realizados mediante SOCIO y Creately por los sujetos experimentales. Las puntuaciones parecen bastante similares para ambos tratamientos.

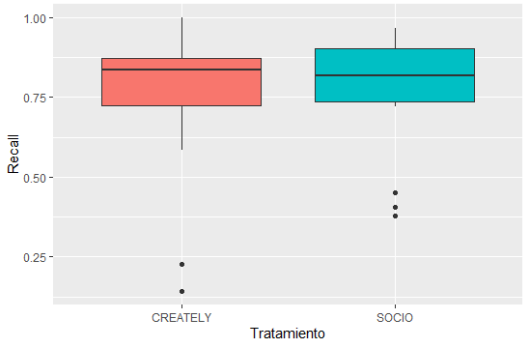


Figura 4.7: *Boxplot* de las valoraciones de recall.

Los resultados del LMM ajustado en el análisis son presentados en la Tabla 4.7. Como podemos apreciar en la Tabla 4.7, el factor estadísticamente significativo es el periodo. Las valoraciones para la primera tarea 1 son superiores que para la segunda, concretamente con una diferencia de 0.17. Por último, d=0.06, SE(d)=0.28, de modo que el tamaño del efecto es pequeño, según [1].

	Estimate	Std.Error	p-value
(Intercept)	0.86	0.07	0
Seq	-0.05	0.09	0.62
Treatment	0.015	0.05	0.75
Period	-0.17	0.05	**0**

Tabla 4.7: LMM de la variable recall.

En definitiva, **se obtienen valoraciones similares de la variable recall para ambos tratamientos.**

Aciertos

En la Figura 4.8 se representa el *boxplot* referente a las valoraciones de aciertos para los diagramas desarrollados mediante SOCIO y Creately por parte de los sujetos experimentales. Como podemos apreciar en dicho diagrama, las puntuaciones para ambos tratamientos parecen muy similares.

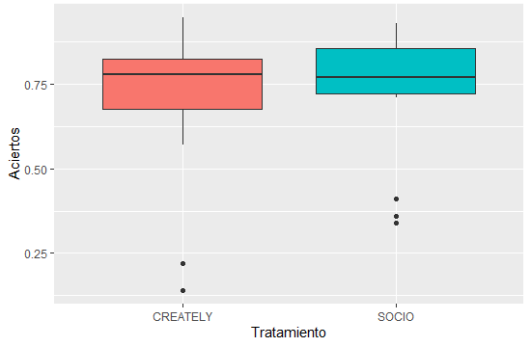


Figura 4.8: *Boxplot* de las valoraciones de aciertos.

Los resultados del LMM ajustado en el análisis son presentados en la Tabla 4.8. Como podemos apreciar en la Tabla 4.8, el factor estadísticamente significativo es el periodo. En la primera tarea se obtiene una valoración superior que para la segunda, con una diferencia de 0.15 puntos. Por último, d=0.13, SE(d)=0.27, lo que nos indica que el tamaño del efecto es pequeño, según [1].

	Estimate	Std.Error	p-value
(Intercept)	0.8	0.07	0
Seq	-0.05	0.09	0.53
Treatment	0.02	0.04	0.53
Period	-0.15	0.04	**0**

Tabla 4.8: LMM de la variable aciertos.

En definitiva, **para la variable aciertos se obtienen resultados similares para los dos tratamientos.**

Error

En la Figura 4.9 se representa el *boxplot* referente a las valoraciones de error con respecto a los diagramas realizados mediante SOCIO y Creately por parte de los sujetos experimentales. Como podemos contemplar, las valoraciones para la variable error aparentan ser mayores para Creately que para SOCIO.

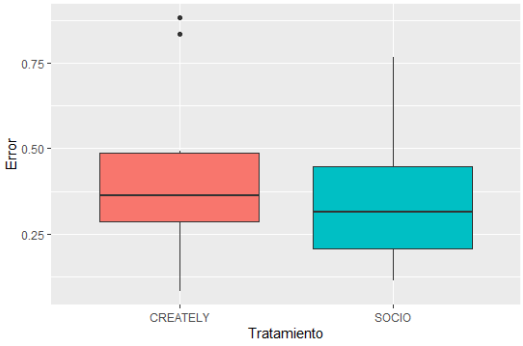


Figura 4.9: *Boxplot* de las valoraciones de error.

Los resultados del LMM ajustado en el análisis son presentados en la Tabla 4.9. Como podemos contemplar en la Tabla 4.9, el factor estadísticamente significativo es el periodo. Las valoraciones de error para la segunda tarea son superiores que para la primera, con una diferencia de 0.19 puntos. Por último, d=-0.1, SE(d)=0.31, de modo que el tamaño del efecto es pequeño, según [1].

	Estimate	Std.Error	p-value
(Intercept)	0.28	0.07	0
Seq	0.029	0.08	0.74
Treatment	-0.023	0.045	0.62
Period	0.19	0.045	**0**

Tabla 4.9: LMM de la variable error.

En definitiva, **parece que con Creately se obtienen más errores que con la herramienta SOCIO.**

4.2. Discusión

En esta sección se lleva a cabo la discusión de las conclusiones obtenidas tras el análisis estadístico de la información recopilada a lo largo del experimento con el fin de confirmar o refutar las hipótesis de la investigación. En la Tabla 4.10 se representa un esquema de las conclusiones obtenidas. Cuando aparezca el símbolo + en las columnas de Tratamiento, Secuencia y Periodo, nos referiremos a que el factor en cuestión es estadísticamente significativo, y cuando aparezca el símbolo / nos referiremos a que no lo es.

Con respecto a la **eficacia**, que se mide a través del grado de completitud de las tareas, el tratamiento no es un factor estadísticamente significativo, esto puede venir causado porque contamos con un tamaño muestral pequeño (16 equipos). Por ahora, como el tratamiento no produce diferencias estadísticamente significativas, no podemos rechazar la hipótesis H.1.0. Sin embargo, el factor periodo/tarea (como mencionamos, se pueden confundir en este estudio experimental) es estadísticamente

significativo, aunque las diferencias que se producen son pequeñas. La tarea 1 obtiene mejores resultados que la tarea 2, es decir, se finaliza con una completitud mayor. Puede ser que la tarea 1 sea más sencilla que la tarea 2, o puede ser por los efectos del *carryover* (que no hayan desaparecido los efectos de aplicar la primera herramienta antes de aplicar la segunda), no podemos afirmarlo debido al pequeño tamaño muestral. Si con un tamaño muestral mayor siguiese habiendo diferencias causadas por las tareas, éstas deberían ser reformuladas.

En cuanto a la característica de la **eficiencia**, que se mide mediante el tiempo y el número de mensajes intercambiados de discusión en la realización de las tareas, no hay ningún factor estadísticamente significativo que afecte al tiempo, sin embargo, los mensajes de discusión vienen nuevamente afectados por el periodo, con pequeñas diferencias, lo que puede ser debido a las mismas causas explicadas en el caso de la eficacia. Con respecto al tratamiento, no se obtienen diferencias significativas ni para el tiempo ni para los mensajes de discusión intercambiados, lo que puede venir provocado por el pequeño tamaño de la muestra. De esta manera, la hipótesis H.2.0 no puede ser rechazada.

De acuerdo a la **satisfacción**, no hay ningún factor estadísticamente significativo que pueda afectar. No se obtienen diferencias significativas que vengan producidas por el factor del tratamiento, esto nuevamente puede venir provocado por el pequeño tamaño de la muestra. De este modo, la hipótesis H.3.0 no puede ser rechazada.

Los resultados para la satisfacción de ambos tratamientos pueden ser contrastados mediante los aspectos tanto positivos como los negativos de dichos tratamientos, los cuales fueron señalados por parte de los usuarios en los cuestionarios de satisfacción, así como las preferencias de cada uno. El 73 % de los participantes indicaron que prefieren SOCIO a Creately.

En cuanto a los aspectos positivos de SOCIO, en general los participantes señalaron la facilidad de uso del chatbot y la rapidez para generar diagramas. Se valora positivamente la integración en las redes sociales, la visualización instantánea del estado del diagrama a través de las imágenes que aporta SOCIO tras cualquier cambio, así como la interacción mediante el lenguaje natural.

En cuanto a Creately, la mayor parte de los participantes señalaron su uso sencillo y su interfaz intuitiva. Valoraron positivamente el poder trabajar de manera colaborativa y respetando el trabajo de los demás participantes del equipo, ya que los elementos que están siendo cambiados se quedan bloqueados. En la secuencia SC-CR los participantes han mostrado menos aspectos positivos que en la secuencia CR-SC, y ciertos criterios que se consideraban positivos en la secuencia CR-SC (intuitiva, facilidad de uso), se consideran negativos en la secuencia SC-CR (poco intuitiva, uso complejo).

De acuerdo a los aspectos negativos de SOCIO, los usuarios destacan que el comando */undo* anula la acción anterior que ha sido realizada por el equipo, pero no la que ha sido llevada a cabo por el usuario que envía dicho comando. Señalan también la dificultad de memorizar los comandos para interactuar con SOCIO, la rigidez en la interpretación del lenguaje y el tratamiento de ambigüedades. Concretamente, varios participantes señalan problemas para asignar el tipo de dato *Date* a un atributo, y sugieren una revisión de este aspecto. También se sugiere una modificación del comando */undo* a fin de poder deshacer una acción propia, avisar de las posibles ambigüedades antes de ejecutar un comando y aportar sugerencias cuando los usuarios se equivocan con los comandos.

Con respecto a los aspectos negativos de Creately, indican que las actualizaciones son lentas y tienen una mala sincronización, ya que en ocasiones no todos los integrantes del equipo visualizaban el mismo diagrama. Asimismo, muchos participantes también señalan la aparición de un mensaje de error durante el uso de la herramienta, la generación de elementos (cuadros de colores) que no pueden ser eliminados e indican la dificultad de modificar la cardinalidad de las relaciones. Sugieren una mejora en la sincronización y corregir los fallos de la aplicación. Además, hay una diferencia notable entre secuencias, ya que los usuarios de la secuencia SC-CR presentan un mayor descontento que los participantes del experimento asignados a la secuencia CR-SC.

Por último, en referencia a la **calidad**, para ninguna métrica (ni accuracy, ni precisión, ni recall, ni aciertos, ni error) el tratamiento es un factor estadísticamente significativo, posiblemente debido al pequeño tamaño muestral. Por tanto, no podemos rechazar la hipótesis H.4.0, por el momento. No obstante, para todas las métricas se obtiene que la tarea/periodo es estadísticamente significativa, aunque con un tamaño de efecto pequeño. Esto puede ser debido a las causas explicadas anteriormente para los casos de la eficacia y la eficiencia.

Variable	Hipótesis	Métrica	Tratamiento	Tamaño del efecto	Secuencia	Periodo/Tarea
Eficacia	H.1.0	Grado de completitud	/	Pequeño	/	+
Eficiencia	H.2.0	Tiempo en finalizar una tarea	/	No hay efecto	/	/
		Nº de mensajes intercambiados de discusión	/	Pequeño	/	+
Satisfacción	H.3.0	Respuestas al cuestionarios SUS	/	Pequeño	/	/
Calidad	H.4.0	Precisión	/	Pequeño	/	+
		Recall	/	Pequeño	/	+
		Accuracy	/	Pequeño	/	+
		Aciertos	/	Pequeño	/	+
		Error	/	Pequeño	/	+

Tabla 4.10: Resumen de las conclusiones del experimento.

5

Agregación de resultados

En este capítulo se realiza el análisis de la agregación formada por una familia de tres experimentos. El primero de ellos [20] con un total de 54 participantes (18 equipos), el segundo [11] se compone de 30 participantes (10 equipos) y finalmente este tercer experimento con un total de 48 participantes (16 equipos). Sobre esta familia formada por 132 participantes (44 equipos) se analiza cada una de las variables respuesta vistas en el capítulo anterior (tiempo, mensajes de discusión, completitud, satisfacción, precisión, recall, accuracy, error y aciertos). En particular, para cada una de ellas se proporcionan:

- Un análisis descriptivo y diagramas de violín divididos por tratamiento (es decir, SOCIO o Creately) y por experimento (Ranci, Andrea o Gemma).

- Un gráfico de perfil en el que se muestra la valoración media de los tratamientos en los estudios experimentales.

Siguiendo las convenciones usadas en otras disciplinas más maduras, como la medicina o la farmacología, se ha ajustado, para analizar esta familia de experimentos, un modelo de regresión lineal (es un modelo IPD *stratified*) de efectos fijos de dos factores: experimento y tratamiento [22]. La elección de dicho modelo de análisis se debe a que está garantizado el acceso a los datos sin procesar de cada uno de los experimentos de la familia, y a que todos los experimentos tienen las mismas variables respuesta. En el modelo establecido, todos los experimentos son replicaciones exactas, las clases de usuarios son parecidas en todos los estudios experimentales y se han realizado pocos experimentos.

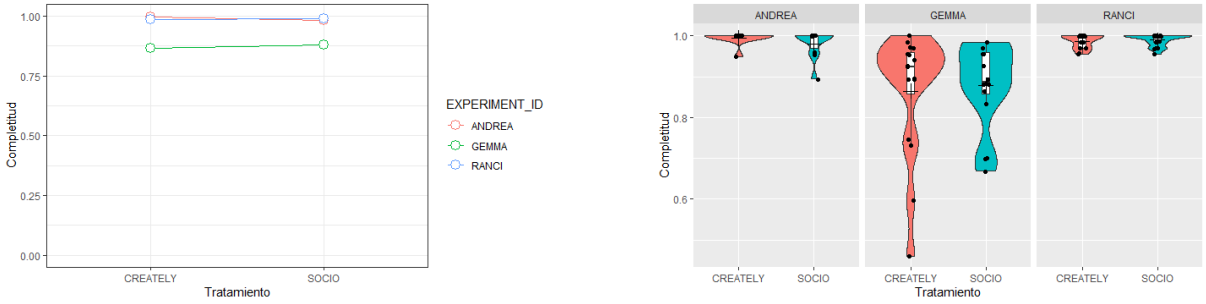
Para poder llevar a cabo el análisis mediante el modelo lineal de efectos fijos debe cumplirse los supuestos de normalidad y homocedasticidad.

El supuesto de normalidad depende del tamaño muestral, en nuestro caso tenemos un tamaño aceptable (44 equipos). Sin embargo, aunque es cierto que bajo el supuesto de normalidad este modelo es más eficiente, el cumplimiento de dicho supuesto no es imprescindible [26] [28]. De todas formas, para poder garantizar este supuesto se ha realizado una comparación de la distribución de los datos a nivel de familia, con la distribución de probabilidad normal. Estos resultados se encuentran en el Apéndice F.

La homocedasticidad se define como la equidad de las varianzas entre los grupos de tratamiento [27]. Los modelos de mínimos cuadrados generalizados [17] que permiten la heterocedasticidad fueron ajustados para comprobar la consistencia de las conclusiones obtenidas a partir de la regresión lineal. Debido a que con ambos modelos (los mínimos cuadrados generalizados y la regresión lineal) se obtuvieron resultados parecidos (similares en tamaño del efecto y significación estadística), se seleccionó el modelo parsimonioso (es decir, el de regresión lineal) para la interpretación de los resultados.

5.1. Eficacia

Las Figuras 5.1(a) y 5.1(b) muestran el gráfico de perfil y el diagrama de violín correspondiente al nivel de completitud de las tareas, respectivamente. En la Tabla 5.1 se representa el esquema de las estadísticas para la completitud, agrupadas por experimento y tratamiento. Como podemos ver en dichos diagramas y en el análisis descriptivo, Creately y SOCIO parecen similares en términos de completitud.



(a) Gráfico de perfil para la completitud.

(b) Diagrama de violín para la completitud.

Figura 5.1: Diagramas para la completitud.

	Tratamiento	Experimento	Equipos	Media	SD	Mediana
1	Creately	Andrea	10	0.99	0.02	1.00
2	SOCIO	Andrea	10	0.98	0.04	1.00
3	Creately	Gemma	16	0.86	0.15	0.92
4	SOCIO	Gemma	16	0.88	0.11	0.89
5	Creately	Ranci	18	0.99	0.02	1.00
6	SOCIO	Ranci	18	0.99	0.01	1.00

Tabla 5.1: Estadísticas para la completitud agrupadas por experimento y tratamiento.

Las Tablas 5.2 y 5.3 representan la tabla ANOVA y el contraste entre tratamientos para la métrica de la completitud.

	numDF	denDF	F-valor	p-valor
(Intercept)	1	42	8282.362	<.0001
Seq	1	40	0.775	0.3841
Treatment	1	42	0.068	0.7955
Period	1	42	3.076	0.0867
Experiment	2	40	14.456	<.0001

Tabla 5.2: Tabla ANOVA para la completitud.

Contrast	Estimate	SE	df	t-ratio	p-valor
Creately-SOCIO	-0.0033	0.0126	42	-0.261	0.7955

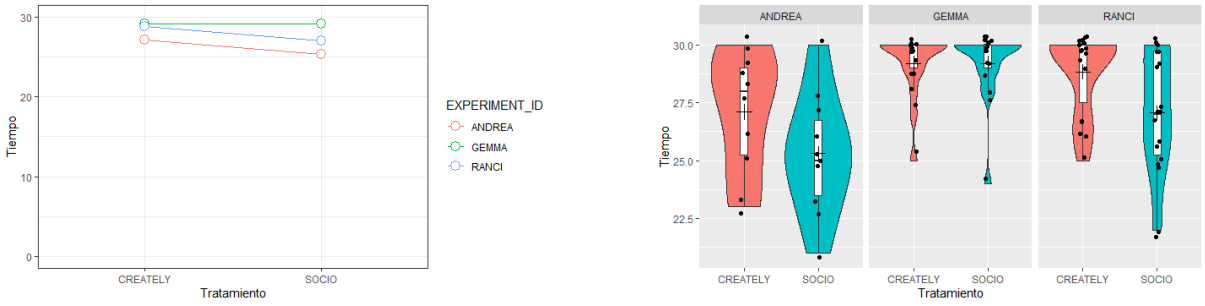
Tabla 5.3: Contraste entre tratamientos para la completitud.

Como podemos ver, hay una pequeña diferencia entre Creately y SOCIO (-0.003), aunque ésta no es estadísticamente significativa. De esta manera, **ambos tratamientos parecen obtener puntuaciones similares en términos de completitud.**

5.2. Eficiencia

Tiempo

Las Figuras 5.2(a) y 5.2(b) muestran el gráfico de perfil y el diagrama de violín para el tiempo empleado en realizar las tareas, respectivamente. En la Tabla 5.4 se representa el esquema de las estadísticas para dicho tiempo, agrupadas por experimento y tratamiento. Como podemos ver en los diagramas y en la tabla de estadísticas descriptivas, el tiempo empleado con SOCIO parece menor que con Creately en dos de nuestros experimentos.



(a) Gráfico de perfil para el tiempo empleado en realizar las tareas.

(b) Diagrama de violín para el tiempo empleado en realizar las tareas.

Figura 5.2: Diagramas para el tiempo empleado en realizar las tareas.

	Tratamiento	Experimento	Equipos	Media	SD	Mediana
1	Creately	Andrea	10	27.1	2.68	28
2	SOCIO	Andrea	10	25.3	2.63	25
3	Creately	Gemma	16	29.19	1.42	30
4	SOCIO	Gemma	16	29.19	1.56	30
5	Creately	Ranci	18	28.83	1.76	30
6	SOCIO	Ranci	18	27.05	2.62	27

Tabla 5.4: Estadísticas para el tiempo agrupadas por experimento y tratamiento.

Las Tablas 5.5 y 5.6 representan la tabla ANOVA y el contraste entre tratamientos para la métrica del tiempo.

	numDF	denDF	F-valor	p-valor
(Intercept)	1	42	15016.244	<.0001
Seq	1	40	0.010	0.9213
Treatment	1	42	6.187	**0.0169**
Period	1	42	0.634	0.4305
Experiment	2	40	11.975	0.0001

Tabla 5.5: Tabla ANOVA para el tiempo.

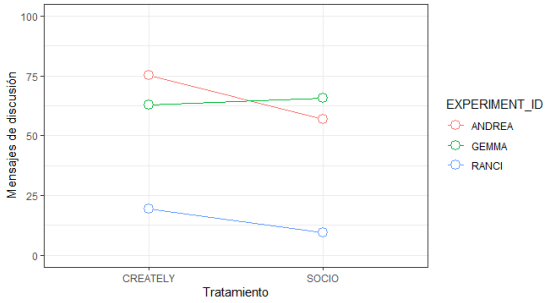
Contrast	Estimate	SE	df	t-ratio	p-valor
Creately-SOCIO	1.14	0.457	42	2.487	**0.0169**

Tabla 5.6: Contraste entre tratamientos para el tiempo.

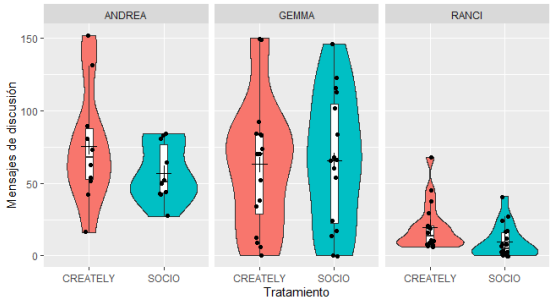
Como podemos ver, la diferencia entre tratamientos es estadísticamente significativa (p-valor = 0.0169). De acuerdo al contraste entre ambos tratamientos, **los participantes tardan de media 1.14 minutos más con Creately que con SOCIO.**

Número de mensajes de discusión

Las Figuras 5.3(a) y 5.3(b) muestran el gráfico de perfil y el diagrama de violín para el número de mensajes intercambiados entre los integrantes de los equipos mientras realizaban las tareas, respectivamente. En la Tabla 5.7 se representa el esquema de las estadísticas para el número de mensajes, agrupadas por experimento y tratamiento. En estos diagramas y estadísticas descriptivas podemos ver que los participantes tienden a enviar un mayor número de mensajes con Creately que con SOCIO, lo que se traduce en un menor esfuerzo a la hora de utilizar la herramienta SOCIO.



(a) Gráfico de perfil para el nº de mensajes de discusión.



(b) Diagrama de violín del nº de mensajes de discusión.

Figura 5.3: Diagramas para los mensajes intercambiados en las tareas.

	Tratamiento	Experimento	Equipos	Media	SD	Mediana
1	Creately	Andrea	10	75.4	40.84	68
2	SOCIO	Andrea	10	57	19.98	51
3	Creately	Gemma	16	63	45.85	70
4	SOCIO	Gemma	16	65.81	46	66
5	Creately	Ranci	18	19.56	16.3	13.5
6	SOCIO	Ranci	18	9.61	11.51	5

Tabla 5.7: Estadísticas del nº de mensajes intercambiados de discusión entre los integrantes de los equipos por experimento y tratamiento.

Las Tablas 5.8 y 5.9 representan la tabla ANOVA y el contraste entre tratamientos para el número de mensajes.

	numDF	denDF	F-valor	p-valor
(Intercept)	1	42	92.38	<.0001
Seq	1	40	0.85428	0.3609
Treatment	1	42	4.118	**0.0488**
Period	1	42	5.696	0.0216
Experiment	2	40	14.442	<.0001

Tabla 5.8: Tabla ANOVA para el nº de mensajes.

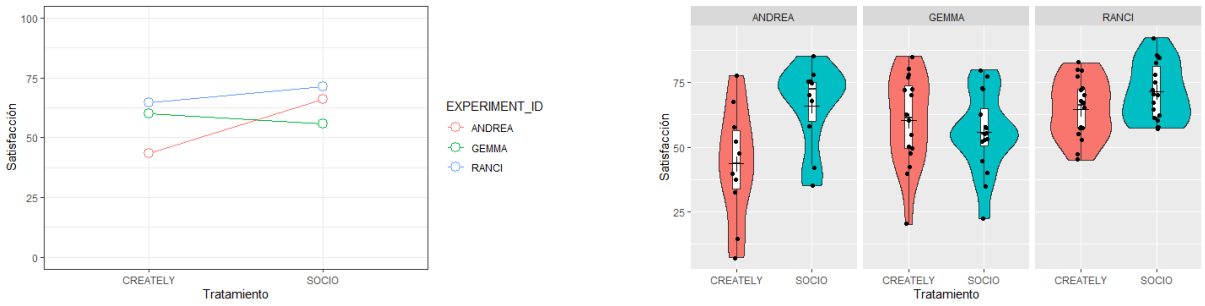
Contrast	Estimate	SE	df	t-ratio	p-valor
Creately-SOCIO	7.23	3.56	42	2.029	**0.0488**

Tabla 5.9: Contraste entre tratamientos del nº de mensajes intercambiados de discusión.

Se observa que la diferencia entre tratamientos de nuevo es estadísticamente significativa (p-valor = 0.0488). En particular, **los participantes tienden a mandar de media 7.23 mensajes más con Creately que con SOCIO.**

5.3. Satisfacción

Las Figuras 5.4(a) y 5.4(b) muestran el gráfico de perfil y el diagrama de violín para la satisfacción, respectivamente. En la Tabla 5.10 se representa el esquema de las estadísticas para dicha métrica, agrupadas por experimento y tratamiento. Como podemos observar, el nivel de satisfacción con SOCIO parece mayor que con Creately en dos de nuestros experimentos, mientras que en el tercero ocurre lo contrario, aunque con una diferencia mucho menor.



(a) Gráfico de perfil para la satisfacción.

(b) Diagrama de violín para la satisfacción.

Figura 5.4: Diagramas para la satisfacción.

	Tratamiento	Experimento	Equipos	Media	SD	Mediana
1	Creately	Andrea	10	43.5	21.86	43.75
2	SOCIO	Andrea	10	66	16.12	72.5
3	Creately	Gemma	16	60.16	17.78	61.25
4	SOCIO	Gemma	16	55.62	15.51	55
5	Creately	Ranci	18	64.72	11.5	66.25
6	SOCIO	Ranci	18	71.32	11.18	70

Tabla 5.10: Estadísticas de la satisfacción agrupadas por experimento y tratamiento.

Las Tablas 5.11 y 5.12 representan la tabla ANOVA y el contraste entre tratamientos para la métrica de la satisfacción.

	numDF	denDF	F-valor	p-valor
(Intercept)	1	42	1350.82	<.0001
Seq	1	40	0.8612	0.359
Treatment	1	42	3.4203	**0.0714**
Period	1	42	5.493	0.0239
Experiment	2	40	5.8296	0.006

Tabla 5.11: Tabla ANOVA para la satisfacción.

Contrast	Estimate	SE	df	t-ratio	p-valor
Creately-SOCIO	-6.16	3.33	42	-1.849	**0.0714**

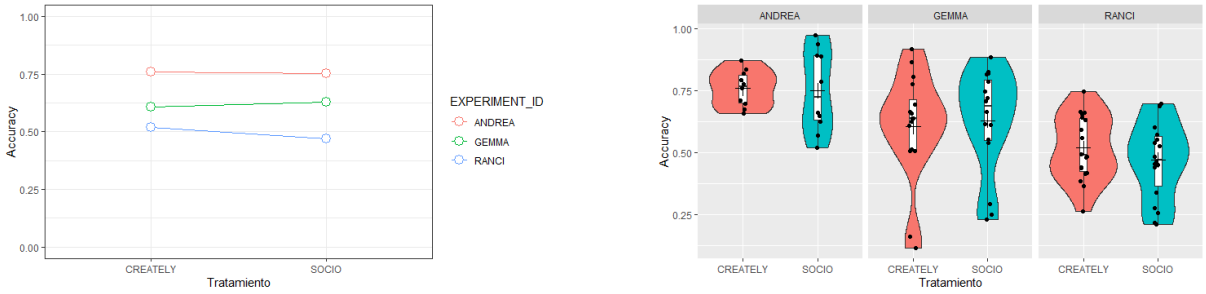
Tabla 5.12: Contraste entre tratamientos para la satisfacción.

La diferencia entre tratamientos para la satisfacción es estadísticamente significativa (p-valor=0.0714) a un nivel de confianza del 0.1, en particular, **los participantes parecen tener mayores puntuaciones de satisfacción con SOCIO que con Creately.**

5.4. Calidad

Accuracy

Las Figuras 5.5(a) y 5.5(b) muestran el gráfico de perfil y el diagrama de violín para las puntuaciones de accuracy, respectivamente. En la Tabla 5.13 se representa el esquema de las estadísticas agrupadas según tratamiento y experimento. Como podemos observar, las valoraciones de accuracy parecen muy similares para ambos tratamientos.



(a) Gráfico de perfil para las puntuaciones de accuracy.

(b) Diagrama de violín para las puntuaciones de accuracy.

Figura 5.5: Diagramas para accuracy.

	Tratamiento	Experimento	Equipos	Media	SD	Mediana
1	Creately	Andrea	10	0.76	0.07	0.77
2	SOCIO	Andrea	10	0.75	0.17	0.73
3	Creately	Gemma	16	0.61	0.22	0.64
4	SOCIO	Gemma	16	0.63	0.21	0.69
5	Creately	Ranci	18	0.52	0.13	0.49
6	SOCIO	Ranci	18	0.47	0.16	0.47

Tabla 5.13: Estadísticas para las puntuaciones de accuracy agrupadas por experimento y tratamiento.

Las Tablas 5.14 y 5.15 representan la tabla ANOVA y el contraste entre tratamientos para la métrica accuracy.

	numDF	denDF	F-valor	p-valor
(Intercept)	1	42	864.8574	<.0001
Seq	1	40	0.2964	0.5892
Treatment	1	42	0.3241	0.5722
Period	1	42	34.5075	<.0001
Experiment	2	40	12.2809	0.0001

Tabla 5.14: Tabla ANOVA para las puntuaciones de accuracy.

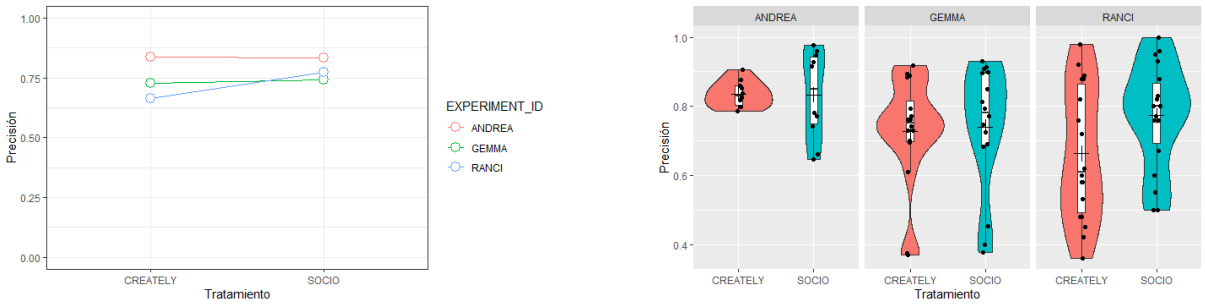
Contrast	Estimate	SE	df	t-ratio	p-valor
Creately-SOCIO	2	40	12.2809	0.0001	0.5722

Tabla 5.15: Contraste entre tratamientos para las puntuaciones de accuracy.

Como hemos visto inicialmente y confirmamos en la tabla ANOVA y el contraste entre tratamiento, **Creately y SOCIO parecen obtener puntuaciones similares de accuracy.**

Precisión

Las Figuras 5.6(a) y 5.6(b) muestran el gráfico de perfil y el diagrama de violín para la precisión, respectivamente. En la Tabla 5.16 se representa el esquema de las estadísticas agrupadas por experimento y tratamiento. Como se aprecia, ambos tratamientos parecen ser similares en términos de la precisión en la mayor parte de los experimentos, aunque en uno de ellos SOCIO obtiene puntuaciones mayores que Creately.



(a) Gráfico de perfil para las puntuaciones de precisión.

(b) Diagrama de violín para las puntuaciones de precisión.

Figura 5.6: Diagramas para la precisión.

	Tratamiento	Experimento	Equipos	Media	SD	Mediana
1	Creately	Andrea	10	0.84	0.04	0.83
2	SOCIO	Andrea	10	0.83	0.13	0.85
3	Creately	Gemma	16	0.73	0.16	0.75
4	SOCIO	Gemma	16	0.74	0.18	0.78
5	Creately	Ranci	18	0.66	0.2	0.61
6	SOCIO	Ranci	18	0.77	0.15	0.8

Tabla 5.16: Estadísticas para las puntuaciones de precisión agrupadas por experimento y tratamiento.

Las Tablas 5.17 y 5.18 representan la tabla ANOVA y el contraste entre tratamientos para la precisión.

	numDF	denDF	F-valor	p-valor
(Intercept)	1	42	1693.41	<.0001
Seq	1	40	0.4214	0.5199
Treatment	1	42	3.9162	**0.0544**
Period	1	42	30.4247	<.0001
Experiment	2	40	3.2065	0.0511

Tabla 5.17: Tabla ANOVA para la precisión.

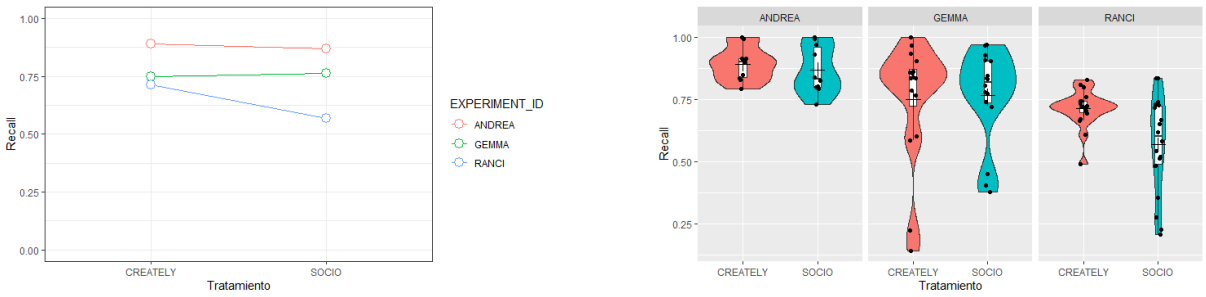
Contrast	Estimate	SE	df	t-ratio	p-valor
Creately-SOCIO	-0.0491	0.0248	42	-1.979	**0.0544**

Tabla 5.18: Contraste entre tratamientos para la precisión.

Como observamos en la tabla ANOVA y el contraste entre tratamientos, la diferencia entre tratamientos es estadísticamente significativa (p-valor=0.0544), y se obtiene una diferencia con una media de 0.0491 puntos a favor de SOCIO. Es decir, **los participantes tienden a obtener mayores puntuaciones de precisión con SOCIO que con Creately.**

Recall

Las Figuras 5.7(a) y 5.7(b) muestran el gráfico de perfil y el diagrama de violín para las puntuaciones de recall, respectivamente. En la Tabla 5.19 se representa el esquema de las estadísticas agrupadas por experimento y tratamiento. Como se puede observar, ambos tratamientos parecen obtener puntuaciones similares de recall en dos nuestros tres experimentos. Sin embargo, en uno de ellos, Creately mejora a SOCIO en términos de la variable recall.



(a) Gráfico de perfil para las puntuaciones de recall.

(b) Diagrama de violín para las puntuaciones de recall.

Figura 5.7: Diagramas para las puntuaciones de recall.

	Tratamiento	Experimento	Equipos	Media	SD	Mediana
1	Creately	Andrea	10	0.89	0.07	0.9
2	SOCIO	Andrea	10	0.87	0.1	0.83
3	Creately	Gemma	16	0.75	0.25	0.84
4	SOCIO	Gemma	16	0.76	0.19	0.82
5	Creately	Ranci	18	0.71	0.08	0.72
6	SOCIO	Ranci	18	0.57	0.2	0.6

Tabla 5.19: Estadísticas para las puntuaciones de recall agrupadas por experimento y tratamiento.

Las Tablas 5.20 y 5.21 representan la tabla ANOVA y el contraste entre tratamientos para la métrica recall.

	numDF	denDF	F-valor	p-valor
(Intercept)	1	42	1234.5463	<.0001
Seq	1	40	0.2032	0.6546
Treatment	1	42	4.5676	**0.0384**
Period	1	42	10.7084	0.0021
Experiment	2	40	9.7432	0.0004

Tabla 5.20: Tabla ANOVA para las puntuaciones de recall.

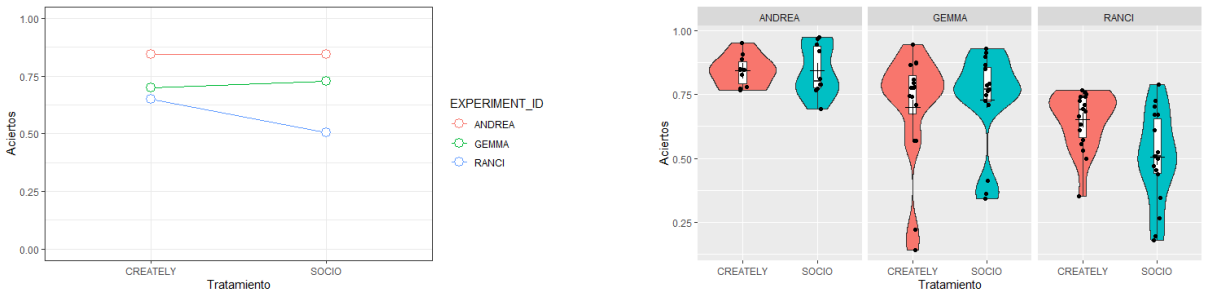
Contrast	Estimate	SE	df	t-ratio	p-valor
Creately-SOCIO	0.0595	0.0279	42	2.137	**0.0384**

Tabla 5.21: Contraste entre tratamientos para las puntuaciones de recall.

Como se observa en la tabla ANOVA y el contraste entre tratamientos, la diferencia de las puntuaciones de recall entre tratamientos es estadísticamente significativa (p-valor=0.0384) con una media de 0.0595 puntos a favor de Creately. Por lo tanto, **Creately supera ligeramente a SOCIO en términos de la variable recall.**

Aciertos

Las Figuras 5.8(a) y 5.8(b) muestran el gráfico de perfil y el diagrama de violín para los aciertos, respectivamente. En la Tabla 5.22 se representa el esquema de las estadísticas para los aciertos agrupadas por experimento y tratamiento. Aunque en dos de nuestros experimentos la puntuación de aciertos es muy similar para ambos tratamientos, en el tercero de los experimentos las puntuaciones para los aciertos con Creately son mucho mayores que con SOCIO.



(a) Gráfico de perfil para las puntuaciones de aciertos.

(b) Diagrama de violín para las puntuaciones de aciertos.

Figura 5.8: Diagramas para las puntuaciones de aciertos.

	Tratamiento	Experimento	Equipos	Media	SD	Mediana
1	Creately	Andrea	10	0.84	0.06	0.85
2	SOCIO	Andrea	10	0.84	0.1	0.8
3	Creately	Gemma	16	0.7	0.23	0.78
4	SOCIO	Gemma	16	0.73	0.19	0.77
5	Creately	Ranci	18	0.65	0.11	0.69
6	SOCIO	Ranci	18	0.5	0.18	0.51

Tabla 5.22: Estadísticas para las puntuaciones de aciertos agrupadas por experimento y tratamiento.

Las Tablas 5.23 y 5.24 representan la tabla ANOVA y el contraste entre tratamientos para la métrica de aciertos.

	numDF	denDF	F-valor	p-valor
(Intercept)	1	42	1139.2168	<.0001
Seq	1	40	0.1317	0.7186
Treatment	1	42	3.7321	**0.0601**
Period	1	42	15.3435	0.0003
Experiment	2	40	13.02	<.0001

Tabla 5.23: Tabla ANOVA para las puntuaciones de aciertos.

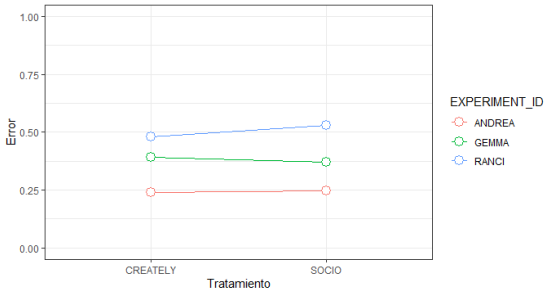
Contrast	Estimate	SE	df	t-ratio	p-valor
Creately-SOCIO	0.0503	0.026	42	1.932	**0.0601**

Tabla 5.24: Contraste entre tratamientos para las puntuaciones de aciertos.

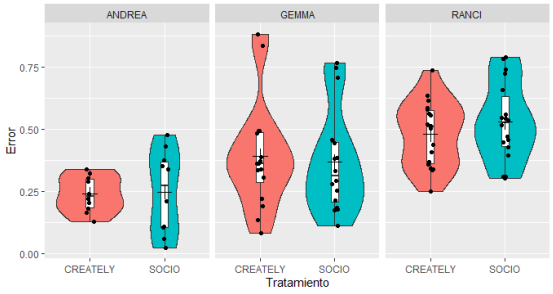
Como se observa en la tabla ANOVA y el contraste, la diferencia entre tratamientos es estadísticamente significativa (p-valor=0.0601) a un nivel de confianza del 0.1, con una media de 0.0503 puntos a favor de Creately. Por tanto, **Creately obtiene una mayor puntuación de aciertos que SOCIO.**

Error

Las Figuras 5.9(a) y 5.9(b) muestran el gráfico de perfil y el diagrama de violín para la puntuación de error, respectivamente. En la Tabla 5.25 se representa el esquema de las estadísticas para los aciertos agrupadas por experimento y tratamiento. Como se aprecia en los diagramas y en las estadísticas descriptivas, parece que ambos tratamientos obtienen puntuaciones similares para la variable de error.



(a) Gráfico de perfil para las puntuaciones de error.



(b) Diagrama de violín para las puntuaciones de error.

Figura 5.9: Diagramas para las puntuaciones de error.

	Tratamiento	Experimento	Equipos	Media	SD	Mediana
1	Creately	Andrea	10	0.24	0.07	0.23
2	SOCIO	Andrea	10	0.25	0.17	0.27
3	Creately	Gemma	16	0.39	0.22	0.36
4	SOCIO	Gemma	16	0.37	0.21	0.31
5	Creately	Ranci	18	0.48	0.13	0.51
6	SOCIO	Ranci	18	0.53	0.16	0.53

Tabla 5.25: Estadísticas para las puntuaciones de error agrupadas por experimento y tratamiento.

Las Tablas 5.26 y 5.27 representan la tabla ANOVA y el contraste entre tratamientos para la métrica de error.

	numDF	denDF	F-valor	p-valor
(Intercept)	1	42	387.6832	<.0001
Seq	1	40	0.2964	0.5892
Treatment	1	42	0.3241	0.5722
Period	1	42	34.5075	<.0001
Experiment	2	40	12.2809	0.0001

Tabla 5.26: Tabla ANOVA para las puntuaciones de error.

Contrast	Estimate	SE	df	t-ratio	p-valor
Creately-SOCIO	-0.0134	0.0236	42	-0.569	0.5722

Tabla 5.27: Contraste entre tratamientos para las puntuaciones de error.

Como se puede observar, **ambos tratamientos tienden a obtener puntuaciones similares de error.**

6

Conclusiones y trabajos futuros

6.1. Conclusiones

Se ha llevado a cabo una réplica de experimentos anteriores [20] [11] con el fin de formar una familia de experimentos y tener un mayor conjunto de datos para obtener unos resultados conjuntos de mayor fiabilidad con respecto a los resultados de cada experimento individualmente.

El método seguido para la evaluación de la usabilidad de SOCIO ha sido compararla con la usabilidad de Creately. En nuestra réplica han participado 48 sujetos, formando un total de 16 equipos de tres miembros cada uno. Tras el experimento se analizaron y discutieron los resultados obtenidos. A continuación, se presentan las conclusiones derivadas:

• El tamaño muestral (16 equipos) para la realización del experimento es pequeño, por lo que no hay diferencias significativas con respecto a la herramienta empleada (SOCIO o Creately) en ninguna métrica asociada a la: calidad, satisfacción, eficacia y eficiencia.

• La tarea parece presentar diferencias significativas estadísticamente en todas las métricas salvo en el tiempo de elaboración de las tareas y en la satisfacción de los usuarios. Concretamente, parecen obtenerse mejores resultados con la tarea 1. Se necesita un tamaño muestral mayor para conocer más concretamente la causa de este hecho.

A nivel de la familia de experimentos, los resultados sugieren lo siguiente:

• Para la **eficacia**, a nivel de familia se obtienen resultados similares con ambos tratamientos. De hecho, como vimos en el Capítulo 5, el tratamiento no produce diferencias estadísticamente significativas, por lo que por el momento no contamos con la suficiente evidencia estadística para poder rechazar la hipótesis H.1.0.

• Los sujetos envían un mayor número de mensajes y tardan más tiempo en completar las tareas con Creately que con SOCIO. Es decir, el chatbot obtiene mejores puntuaciones en términos de la **eficiencia**. En particular, el tratamiento es un factor estadísticamente significativo en ambas métricas, por lo que se rechaza la hipótesis H.2.0.

• En cuanto a la **satisfacción**, los participantes están más satisfechos con SOCIO. Como vimos, el factor tratamiento presenta diferencias estadísticamente significativas con respecto a la satisfacción de los usuarios, por lo que se rechaza la hipótesis H.3.0.

• Finalmente, atendiendo a la **calidad** de los diagramas, mediante SOCIO se obtienen mayores puntuaciones de precisión, mientras que Creately parece mejor en términos de las variables recall y aciertos. De hecho, como vimos anteriormente, el tratamiento genera diferencias significativas en estas tres métricas, de modo que la hipótesis H.4.0 es aceptada parcialmente (como precisión a favor

de SOCIO y recall y aciertos a favor de Creately). En otras palabras, parece que con SOCIO los participantes crean un menor número de clases, y la mayoría de esas clases son correctas ya que forman parte de la solución ideal. Sin embargo, los participantes crean más clases con Creately y tienen la impresión de conseguir mejores resultados con Creately que con SOCIO.

Como podemos comprobar, gracias al aumento del tamaño muestral en la familia y su consecuente aumento en la potencia estadística de los resultados, la información proporcionada a nivel de familia es mucho más precisa que la información de cada uno de los experimentos individualmente, ya que en muchos casos no se han observado diferencias significativas, mientras que realmente sí que las hay.

De esta manera, finalmente hemos concluido que la usabilidad de SOCIO posee un efecto ventajoso en la gran mayoría de aspectos, en comparación con la aplicación web Creately. Dichos aspectos son la eficiencia, por el tiempo de elaboración de las tareas y la cantidad de mensajes intercambiados de discusión durante las mismas, la satisfacción y la precisión de los modelos con respecto a la calidad.

6.2. Trabajos futuros

Esta investigación contribuye a los estudios empíricos de la evaluación de la usabilidad de los chatbots y, en particular, del chatbot SOCIO. Las diferencias significativas estadísticamente pero con un pequeño tamaño del efecto suponen un motivo para continuar con la investigación en las líneas siguientes:

- Realizar más réplicas de este estudio experimental, pero con una clase de usuarios diferentes a los que han participado en esta familia. Es decir, profesionales de la informática en lugar de estudiantes.

- Replantear el diseño experimental aumentando el tiempo máximo permitido de realización de las tareas, ya que no hay diferencia en la velocidad en dos de nuestros tres experimentos.

- Cambiar el diseño de las tareas para comprobar si está mediando en las relaciones analizadas.

- Ampliar la familia de experimentos con ese cambio para comprobar la consistencia de los resultados.

Bibliografía

[1] M. Borenstein, L.V. Hedges, J.P. Higgins, and H.R. Rothstein. *Introduction to Meta-Analysis*. John Wiley and Sons, 2011.

[2] J. Brooke. SUS-A quick and dirty usability scale. *Usability Evaluation in Industry*, 189(194):4–7, 1996.

[3] M.L. Chen, and H.C. Wang. How personal experience and technical knowledge affect using conversational agents. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, pages 53–58, 2018.

[4] A. Cheng, V. Raghavaraju, J. Kanugo, Y.P. Handrianto, and Y. Shang. Development and evaluation of a healthy coping voice interface application using the Google Home for elderly patients with type 2 diabetes. In *Proceedings of the 15th IEEE Annual Consumer Communications and Networking Conference*, pages 1–5, 2018.

[5] A. Field, J. Miles, and Z. Field. *Discovering Statistics Using R*. Sage Publications, 2012.

[6] J.P. Higgins, and S. Green. *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley and Sons, 2011.

[7] ISO Std. ISO 9241-11. Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)-Part II: Guidance on Usability. *International Organization for Standardization*, 1998.

[8] ISO/IEC Std. ISO/IEC 25010. Systems and Software Engineering. System and Software Product. Quality Requirements and Evaluation (SQuaRE). System and Software Quality Models. *International Organization for Standardization*, 2010.

[9] M. Jain, R. Kota, P. Kumar, and S.N. Patel. Convey: Exploring the use of a context view for chatbots. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 468–477. ACM, 2018.

[10] I. Lopatovska, K. Rink, I. Knight, K. Raines, K. Cosenza, H. Williams, P. Sorsche, D. Hirsch, Q. Li, and A. Martinez. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science*, (4):1–14, 2018.

[11] A. Nevado Labrador. Evaluación Empírica de la Usabilidad de un Chatbot. Trabajo Fin de Grado. Directora: S.T. Acuña. Doble Grado en Ingeniería Informática y Matemáticas, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 2019.

[12] Q.N. Nguyen, and A.C. Sidorova. Understanding user interactions with a chatbot: A selfdetermination theory approach. In *Proceeding of the 24th Americas Conference on Information Systems 2018: Digital Disruption*, pages 1–5, 2018.

[13] J.I. Panach Navarrete, O. Dieste, B. Marin, S. Espana, S. Vegas, O. Pastor, and N. Juristo. Evaluating model-driven development claims with respect to quality: A family of experiments. *IEEE Transactions on Software Engineering*, pages 1–18, 2018.

[14] J. Pérez, Y. Sánchez, F.J. Serón, and E. Cerezo. Interacting with a semantic affective ECA. In J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, I. Leite, and S. Kopp, editors, *Intelligent Virtual Agents*, pages 374–384. Springer, 2017.

[15] S. Perez-Soler, E. Guerra, and J. de Lara. Collaborative modeling and group decision making using chatbots in social networks. *IEEE Software*, 35(6):48–54, 2018.

[16] S. Pérez-Soler, E. Guerra, J. de Lara, and F. Jurado. The rise of the (modelling) bots: Towards assisted modelling via social networks. In *Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering*, pages 723–728. IEEE Press, 2017.

[17] J. Pinheiro, and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer Science and Business Media, 2006.

[18] R. Ren. An Experimental Study on the Usability of Chatbots. Trabajo Fin de Máster. Directora: S.T. Acuña, and J.W. Castro. Máster en Investigación e Innovación en Inteligencia Computacional y Sistemas Interactivos, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 2019.

[19] R. Ren, J.W. Castro, S.T. Acuña, and J. de Lara. Usability of chatbots: A systematic mapping study. In *Proceedings of the 31st International Conference on Software Engineering and Knowledge Engineering*, pages 479–484, 2019.

[20] R. Ren, J.W. Castro, A. Santos, S. Pérez-Soler, S.T. Acuña, and J. de Lara. Collaborative modelling: Chatbots or on-line tools? An experimental study. In *Proceedings of the Evaluation and Assessment in Software Engineering*, pages 260–269, 2020.

[21] A. Santos, O.S. Gómez, and N. Juristo. Analyzing families of experiments in SE: A systematic mapping study. *IEEE Transactions on Software Engineering*, 46(5):566–583, 2020.

[22] A. Santos, S. Vegas, M. Oivo, and N. Juristo. A procedure and guidelines for analyzing groups of software engineering replications. *IEEE Transactions on Software Engineering*, pages 1–22, 2019.

[23] C. Sinoo, S. van der Pal, O.A.B. Henkemans, A. Keizer, B.P. Bierman, R. Looije, and M.A. Neerincx. Friendship with a robot: Children's perception of similarity between a robot's physical and virtual embodiment that supports diabetes self-management. *Patient Education and Counseling*, 101(7):1248–1255, 2018.

[24] M.L. Tielman, M.A. Neerincx, R. Bidarra, B. Kybartas, and W.P. Brinkman. A therapy system for post-traumatic stress disorder using a virtual agent and virtual storytelling to reconstruct traumatic memories. *Journal of Medical Systems*, 41(8):125, 2017.

[25] S. Vegas, C. Apa, and N. Juristo. Crossover designs in software engineering experiments: Benefits and perils. *IEEE Transactions on Software Engineering*, 42(2):120–135, 2016.

[26] A.J. Vickers. Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Medical Research Methodology*, 5(35):1–12, 2005.

[27] A. Whitehead. *Meta-Analysis of Controlled Clinical Trials*. John Wiley and Sons, 2002.

[28] J.C.F. Winter. Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, 18(1):1–13, 2013.

Glosario

Agregación Combinación de los resultados obtenidos de una serie de estudios experimentales para analizar el desempeño de dos tratamientos, en un contexto dado, con el fin obtener un resultado final único.

Carryover Es la persistencia del efecto de un tratamiento cuando se aplica otro tratamiento diferente.

Creately Herramienta web empleada en el experimento que posibilita el desarrollo de diferentes tipos de diagramas colaborativamente, entre otros se encuentran los diagramas de clases.

Crossover Caso particular de diseño intra-sujetos donde distintos sujetos experimentales aplican diferentes tratamientos que serán evaluados en distinto orden.

Experimento verdadero Diseño experimental en el que los sujetos experimentales se asignan de manera aleatoria a los tratamientos a evaluar.

Meta-análisis Instrumentos estadísticos adecuados para combinar los resultados de un conjunto de experimentos singulares que forman una familia de experimentos.

Modelo de efectos aleatorios Método estadístico donde se parte de la base de que los estudios singulares estiman a una distribución de tamaños del efecto.

Modelo de efectos fijos Método estadístico donde se asume que cada uno de los estudios singulares del meta-análisis están estimando a un mismo (y único) tamaño del efecto.

Réplica Repetición del experimento con una configuración idéntica de los factores del mismo.

SOCIO Chatbot, el cual asiste en el diseño de diagramas de dominio interpretando mensajes en lenguaje natural en inglés.

Within-subject Tipo de diseño donde cada sujeto experimental aplica todos los tratamientos que se van a evaluar.

Apéndices

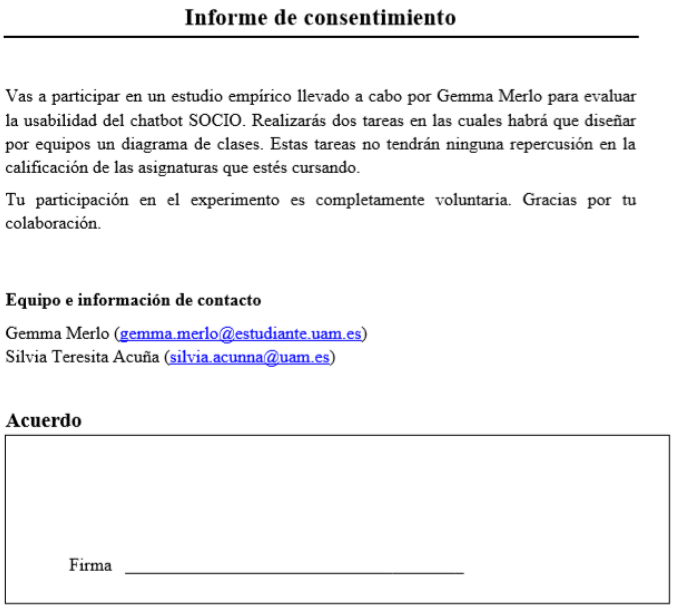
A

Documentos del experimento

En este apéndice se encuentran los documentos utilizados durante el experimento. En el comienzo de cada una de las sesiones, los participantes debían rellenar un documento de consentimiento informado y rellenar un cuestionario para recoger información básica acerca de los sujetos. Para realizar las tareas, los enunciados de las mismas eran entregados y, al finalizar cada una de las tareas, el cuestionario de satisfacción acerca de la herramienta que acababan de emplear para realizar la tarea (SOCIO o Creately).

A.1. Documentos iniciales

En la Figura A.1 podemos observar el informe de consentimiento que debían firmar los participantes al comenzar las sesiones, y en la Figura A.2 se muestra el cuestionario de familiaridad que se les entregaba después.



Informe de consentimiento

Vas a participar en un estudio empírico llevado a cabo por Gemma Merlo para evaluar la usabilidad del chatbot SOCIO. Realizarás dos tareas en las cuales habrá que diseñar por equipos un diagrama de clases. Estas tareas no tendrán ninguna repercusión en la calificación de las asignaturas que estés cursando.

Tu participación en el experimento es completamente voluntaria. Gracias por tu colaboración.

Equipo e información de contacto

Gemma Merlo (gemma.merlo@estudiante.uam.es)
Silvia Teresita Acuña (silvia.acunna@uam.es)

Acuerdo

Firma _____

Figura A.1: Informe de consentimiento.

GRUPO ___

Cuestiones generales: Para cada una de las siguientes cuestiones rellena o marca la casilla correspondiente.

Edad. . . . . ☐

Sexo. . . . . Hombre ☐ Mujer ☐

¿Eres estudiante o graduado en informática? . . . . . Sí ☐ No ☐

¿Has utilizado alguna vez Telegram?. . . . . Sí ☐ No ☐

¿Has utilizado alguna vez un chatbot?. . . . . Sí ☐ No ☐

¿Qué redes sociales sueles utilizar?. . . WhatsApp ☐ Telegram ☐ Twitter ☐ Facebook ☐ Instagram ☐

Puntúa tu grado de uso de las redes sociales (1-poco/ninguno, 5-intensivo). 1 2 3 4 5

Puntúa tu grado de uso de Telegram (1-poco/ninguno, 5-intensivo). 1 2 3 4 5

Puntúa tu nivel de inglés (1-novato, 5-experto). 1 2 3 4 5

Puntúa tu grado de conocimiento sobre diagramas de clases (1-novato, 5-experto). 1 2 3 4 5

Puntúa tu nivel de conocimiento sobre chatbots (1-novato, 5-experto). 1 2 3 4 5

Puntúa tu grado de uso de chatbots (1-poco/ninguno, 5-intensivo). 1 2 3 4 5

Versión de Telegram empleada durante la sesión: . . . . . SmartPhone o Tablet ☐ Web ☐ Escritorio ☐

Figura A.2: Cuestionario de familiaridad.

A.2. Enunciados y soluciones

En las Figuras A.3 y A.4 se muestran los enunciados para que los equipos desarrollen los diagramas de clases en las dos tareas.

TAREA 1

A shop requests an application to manage their products and their clients. They have three types of products: clothes, shoes and bags. All products have an identifier, a name, a color, a description, a price and a category. In some seasons, products may have a discount. The clothes and shoes have a size, and the shoes can be of different heights. The shop wants to visualize all this information about their products, and also, a photo and the number of units.

The shop has the name, address and telephone number of its clients. Each client has an identifier. Clients can place orders. The shop wants to be able to register the orders of each client in the application, in order to see the date on which the order will be made, its identifier and the products it contains.

Figura A.3: Enunciado de la primera tarea.

TAREA 2

A school, whose name and address are known, requests an application to organize its teachers, students and subjects. The school teaches different subjects depending on the academic year. Each subject has several lessons that can be managed from the application. Exams are performed to evaluate each subject. The school wants to be able to specify the questions, the date and the weight of the exams in the subject through the application. Several classes are taught per subject, in a specific classroom, at a specific day and time. Each class has several students and is given by a single teacher. The school has the full name, address, telephone number and date of its teachers and students. In addition, every person belonging to the school has an identifier.

Figura A.4: Enunciado de la segunda tarea.

Las Figuras A.5 y A.6 muestran las soluciones ideales de las tareas 1 y 2, respectivamente. Las soluciones ideales son tomadas como referentes para valorar los diagramas diseñados durante el experimento por parte de los sujetos.

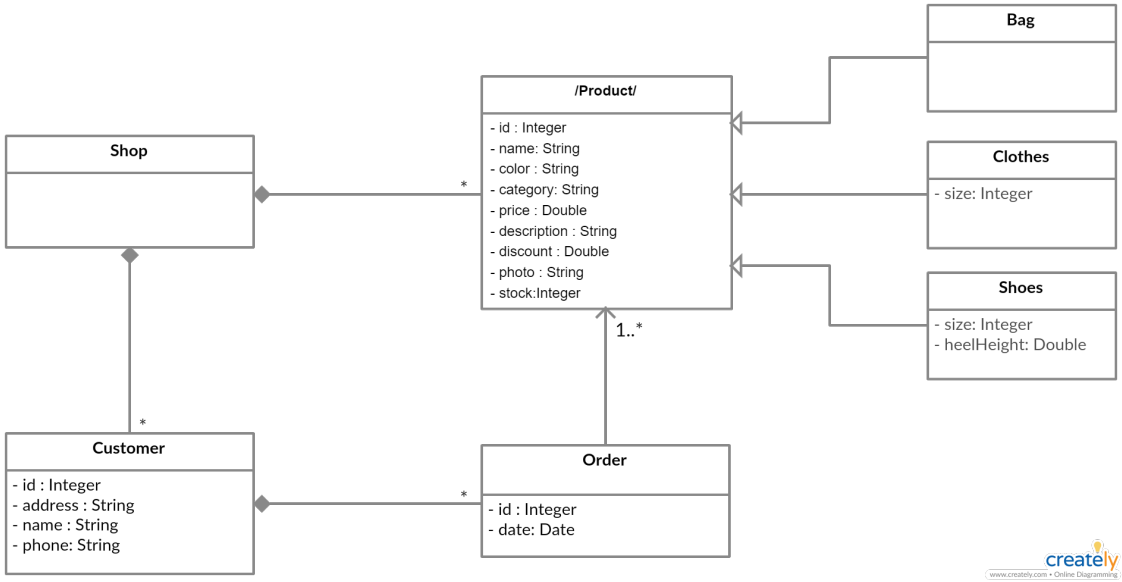


Figura A.5: Diagrama ideal para la primera tarea.

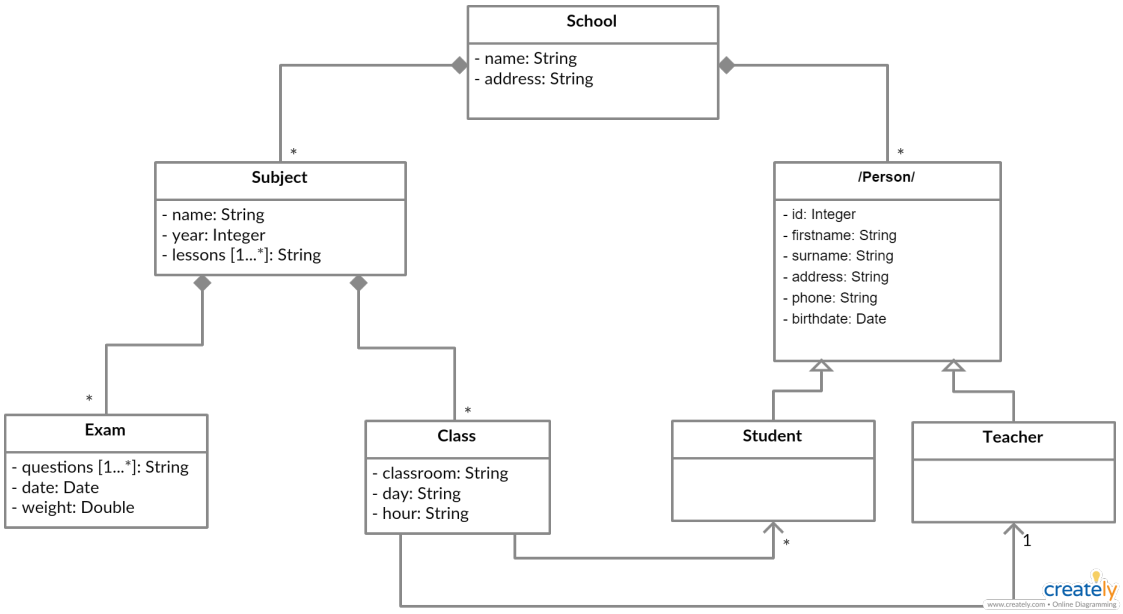


Figura A.6: Diagrama ideal para la segunda tarea.

A.3. Cuestionarios del experimento

En la Figura A.7 se observa el cuestionario SUS (de satisfacción) adaptado en [16] y cumplimentado por los usuarios tras la realización de cada tarea del experimento. El cuestionario SUS entregado a los participantes después de la última tarea incluye además un apartado acerca de la preferencia del usuario en la que éste debe elegir una de las dos herramientas (SOCIO o Creately).

GRUPO ___

HERRAMIENTA _____

Instrucciones: Para las siguientes afirmaciones, marca la casilla que mejor describa tus reacciones a la herramienta.

totalmente de acuerdo ➡
⬅ totalmente en desacuerdo

Creo que me gustaría usar esta herramienta con frecuencia.	1 2 3 4 5
Encontré esta herramienta innecesariamente compleja.	1 2 3 4 5
Creo que la herramienta es fácil de usar.	1 2 3 4 5
Creo que necesitaría ayuda para poder usar esta herramienta.	1 2 3 4 5
He encontrado que las diversas funciones de esta herramienta estaban bien integradas.	1 2 3 4 5
Creo que hay demasiadas funciones inconsistentes en esta herramienta.	1 2 3 4 5
Creo que la mayoría de la gente puede aprender a usar esta herramienta muy rápidamente.	1 2 3 4 5
He encontrado esta herramienta muy engorrosa/incómoda de usar.	1 2 3 4 5
Me sentí muy seguro de lo que hacía al usar esta herramienta.	1 2 3 4 5
Tengo que aprender un montón de cosas antes de poder usar esta herramienta.	1 2 3 4 5

Por favor, indica tres aspectos positivos que quieras resaltar sobre la herramienta:

_____
_____
_____

Por favor, indica tres aspectos negativos de la herramienta:

_____
_____
_____

¿Tienes alguna sugerencia de mejora?:

_____

Figura A.7: Cuestionario SUS.

B

Herramientas del experimento

B.1. Creately

Creately es una herramienta colaborativa para la elaboración de diversos tipos de diagramas, entre ellos se encuentran los diagramas de clases. En la Figura B.1 se muestra la apariencia de la aplicación. En el menú lateral izquierdo se pueden seleccionar los elementos necesarios para la elaboración del diagrama, y arrastrarlos al cuadro central para utilizarlos.

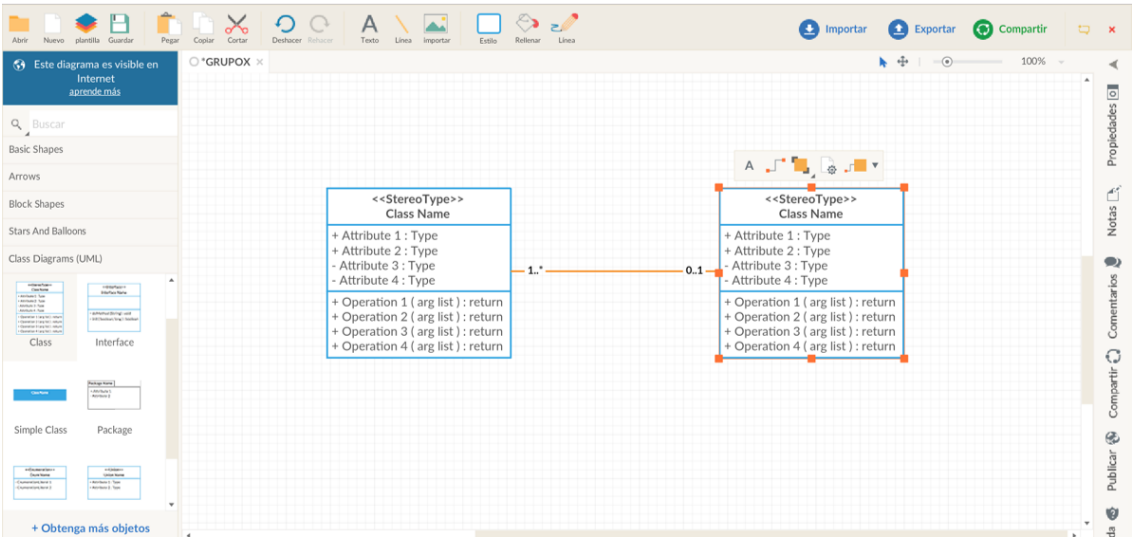


Figura B.1: Interfaz de Creately.

Cuando se selecciona una clase se muestra el menú de la misma, en el que podremos seleccionar relaciones y editar las clases, como podemos ver en la Figura B.2. Del mismo modo, al seleccionar una relación aparece el menú correspondiente a dicha relación mediante el cual podremos determinar el tipo de relación que se quiera, como observamos en la Figura B.3.

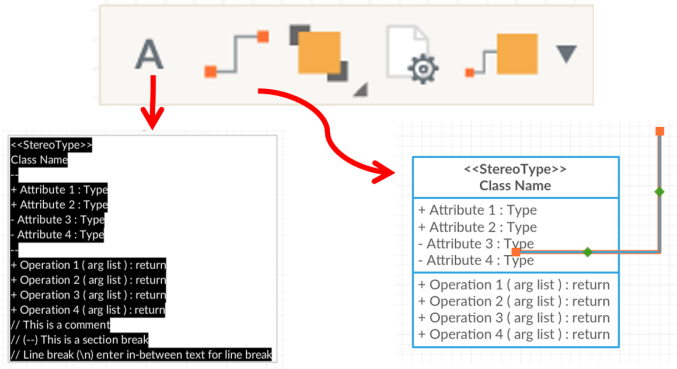


Figura B.2: Menú que se refiere a la clase.

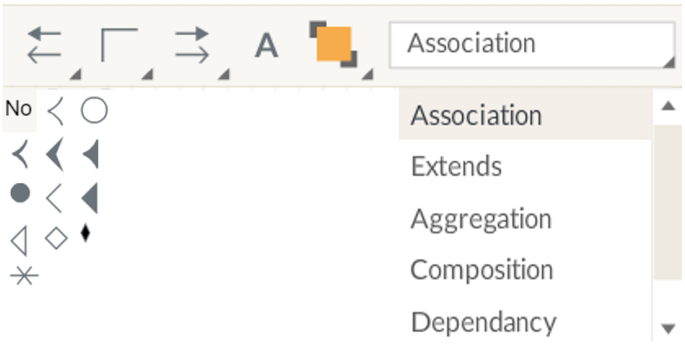


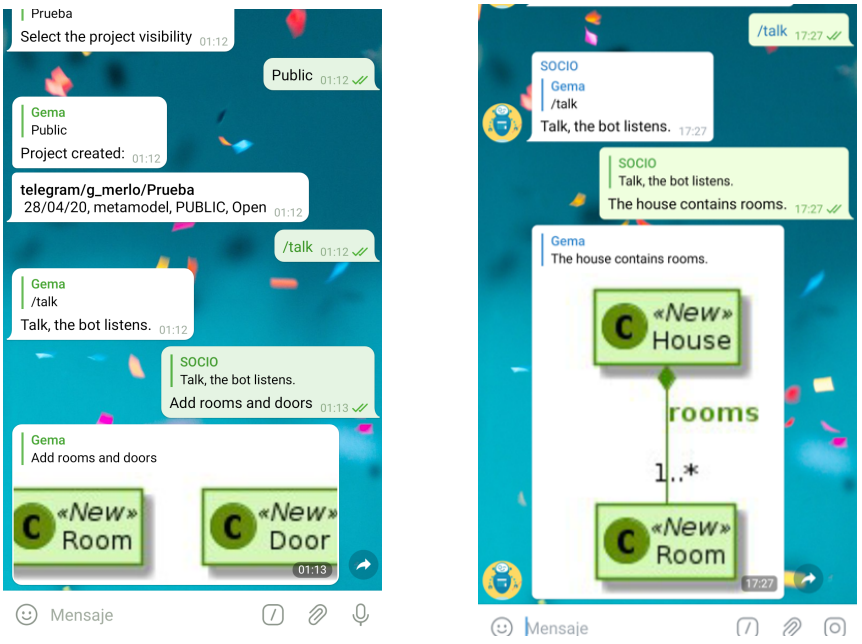
Figura B.3: Menú que se refiere a la relación.

B.2. Tipos de mensajes dirigidos a SOCIO

En la Tabla B.1 se muestran los comandos necesarios para interaccionar con SOCIO. Dichos comandos presentan una sintaxis flexible, como se muestra en la Figura B.4(a). Además, en la Figura B.4(b) se representa un ejemplo de mensaje de naturaleza descriptiva.

Comando	Descripción
\branch	Creación de una nueva rama para el proyecto.
\delproject	Eliminar un proyecto.
\get	Enviar un archivo con el modelo.
\help	Vínculo a la web `https://saraperezsoler.github.io/ModellingBot/`.
\history	Estadísticas e histórico de mensajes.
\newproject	Creación de un proyecto nuevo.
\projectmanager	Gestión de usuarios y visibilidad de un proyecto.
\projects	Lista con los existentes proyectos.
\redo	Rehacer la última acción.
\setproject	Seleccionar un proyecto para trabajar sobre él.
\show	Mostrar el estado actual del modelo.
\start	Mostrar la totalidad de los comandos.
\talk	Para enviar mensajes a SOCIO con el fin de realizar el modelo. Pueden ser mensajes descriptivos o comandos (órdenes).
\undo	Deshacer la última acción.
\validate	Validar el modelo.

Tabla B.1: Comandos para interaccionar con SOCIO.



(a) Procesamiento de comandos.

(b) Procesamiento de mensajes descriptivos.

Figura B.4: Procesamiento de distintos tipos de mensajes.

C

Evaluación de la calidad y la completitud

C.1. Evaluación de la calidad

Este apéndice detalla el modo de proceder para la evaluación de la calidad de los diagramas desarrollados por los sujetos experimentales a lo largo de las tareas del experimento.

Para la evaluación de la calidad [11] se ha comparado el diagrama elaborado por cada uno de los equipos con la solución ideal de la tarea en cuestión. Las soluciones ideales de cada una de las tareas se muestran en el Apéndice A. Dicha evaluación se lleva a cabo a través de la matriz de confusión de la Figura C.1.

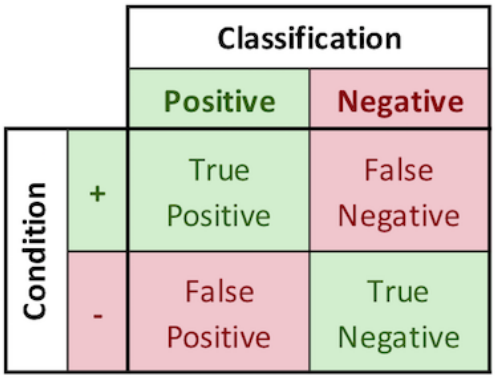


Figura C.1: Matriz de confusión para evaluar la calidad.

Donde “Condition” hace referencia a la solución ideal, y “Classification” al diagrama elaborado por cada uno de los equipos durante las tareas. Los elementos de la matriz se detallan a continuación:

- *True positives* (TP): cantidad de componentes que están tanto en la solución ideal como en el diagrama desarrollado por un sujeto experimental.

- *False positives* (FP): cantidad de componentes que están en el diagrama desarrollado por un sujeto experimental, pero no se encuentran en la solución ideal.

- *False negatives* (FN): cantidad de componentes que no están en el diagrama desarrollado por un sujeto experimental, y sí se encuentran en la solución ideal.

- *True negatives* (TN): Su valor siempre es nulo, en el caso de la comparación de diagramas.

Cada una de las componentes de un diagrama está constituida por varias partes. En la Tabla C.1 se muestran las partes correspondientes a cada componente y el sistema de valoración utilizado.

Componente	**Clase**			**Relación**			**Atributo**	
Partes	Presencia	Nombre	Tipo	Presencia	Tipo	Cardinalidad	Presencia	Tipo
Valoración	0.65	0.1	0.25	0.5	0.25	0.25	0.75	0.25

Tabla C.1: Método de valoración para las componentes de un diagrama.

De esta manera, se trata de sumar cada una de las partes correctas de cada componente, y si la totalidad de dicha componente es correcta, su valor cuenta como 1 punto.

Si una parte de una componente es incorrecta (como por ejemplo, una clase que se considera abstracta cuando en realidad no lo es), la valoración correspondiente a dicha parte se cuenta como FP.

Si a una componente le falta alguna parte (como por ejemplo, si no se indica el tipo de un atributo), la puntuación correspondiente a dicha parte se cuenta como FN.

Los valores de *true positives*, *false positives*, *false negatives* y *true negatives* se emplean para el cálculo de las métricas de la calidad, como se ha indicado anteriormente:

- Precisión = $\frac{TP}{TP+FP}$.
- Recall = $\frac{TP}{TP+FN}$.
- Accuracy = $\frac{TN+TP}{TP+TN+FP+FN}$.
- Aciertos = $\frac{TP}{N^{\circ}\ elementos\ diagrama\ ideal}$.
- Error = $\frac{FP+FN}{TP+TN+FP+FN}$.

La métrica de la precisión indica el porcentaje de elementos correctos en el diagrama elaborado por un equipo, en función de los elementos del diagrama ideal. Recall representa la proporción de elementos del diagrama ideal presentes en el diagrama desarrollado por un sujeto experimental. La métrica accuracy es una combinación de precisión y recall. La métrica de error representa cuántos elementos faltan en el diagrama elaborado por un equipo. Finalmente, la métrica de aciertos es la tasa de éxito de cada equipo, en comparación con la solución ideal.

C.2. Evaluación de la completitud

Este Apéndice detalla el modo de proceder para evaluar el nivel de completitud de los diagramas desarrollados por los sujetos experimentales a lo largo de las tareas. Para la evaluación del grado de completitud se ha comparado el diagrama elaborado por cada uno de los equipos con la solución ideal de la tarea en cuestión (mostradas en el Apéndice A).

Si un equipo ha completado una tarea, el valor máximo para la completitud es 1. En caso de que la tarea no haya sido terminada, cada elemento que falte en el diagrama del equipo cuenta 0.03, salvo el tipo del atributo, que cuenta 0.015. De esta manera, se puntúa la totalidad de elementos que faltan en el diagrama, siendo el grado de completitud, la diferencia entre el valor máximo (1) y la puntuación de los elementos de los que carece el diagrama del equipo.

D

Análisis por secuencias y tareas

Este anexo presenta los *boxplots* de las mediciones de eficacia, eficiencia, y satisfacción como características de usabilidad, al igual que las correspondientes a la calidad de los modelos creados a lo largo de las tareas realizadas por los equipos en el presente experimento. Estos *boxplots* muestran los datos para cada una de las métricas, o bien en función del tratamiento (SOCIO o Creately) junto con la secuencia SOCIO-Creately (SC-CR) o la secuencia Creately-SOCIO (CR-SC), o bien en función del tratamiento y del periodo o la tarea (la realizada en primer lugar o en segundo lugar).

D.1. Eficacia

La forma para medir la eficacia es a través del nivel de completitud de realización de las tareas por parte de los equipos. En la Figura D.1 se representa el *boxplot* correspondiente a dicho nivel de completitud según tratamiento-secuencia. La secuencia CR-SC muestra que la aplicación web Creately obtiene una puntuación de completitud ligeramente mayor que la del chatbot SOCIO, en cambio la secuencia SC-CR obtiene puntuaciones muy similares para ambos tratamientos, aunque Creately presenta valores más bajos.

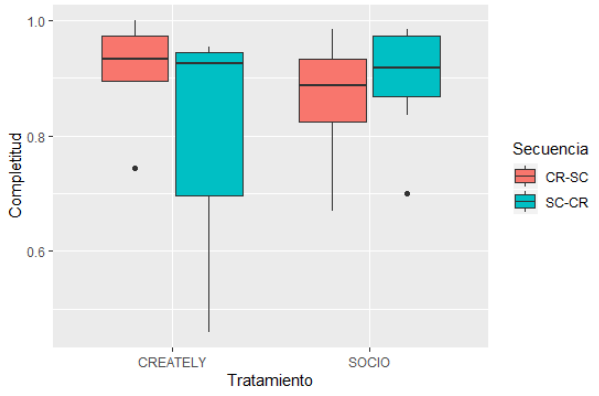


Figura D.1: *Boxplot* de la eficacia según tratamiento-secuencia.

En la Figura D.2 se muestra el *boxplot* del nivel referido a la completitud en el que los equipos terminaron las tareas en relación con el tratamiento-periodo. Los datos presentan similitud para ambos tratamientos, con una mediana ligeramente superior para Creately en la tarea 2, aunque en Creately se observan valores más bajos en la tarea 2.

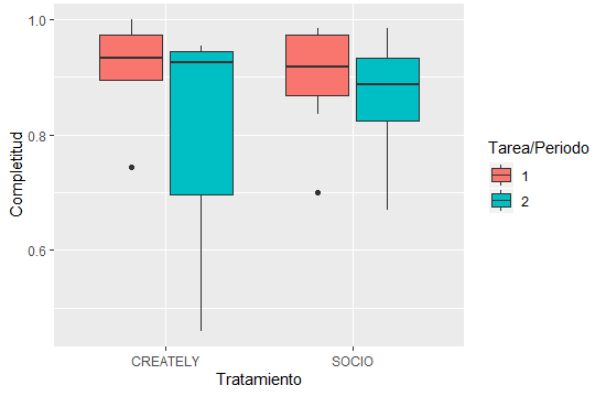


Figura D.2: *Boxplot* de la eficacia según tratamiento-periodo.

D.2. Eficiencia

Como se ha mencionado, las métricas utilizadas para medir la eficiencia son el tiempo utilizado por los sujetos a fin de realizar las tareas y el esfuerzo de interacción representado por el nº de mensajes de discusión generados a lo largo de dichas tareas. A continuación, se presentan los *boxplots* de dichas métricas según tratamiento-secuencia, y según tratamiento-tarea (-periodo).

Tiempo

En la Figura D.3 se representa el *boxplot* referente al tiempo utilizado para realizar las tareas por parte de los equipos según tratamiento-secuencia. Como podemos observar, los datos son muy similares para ambos tratamientos: secuencia CR-SC y secuencia SC-CR.

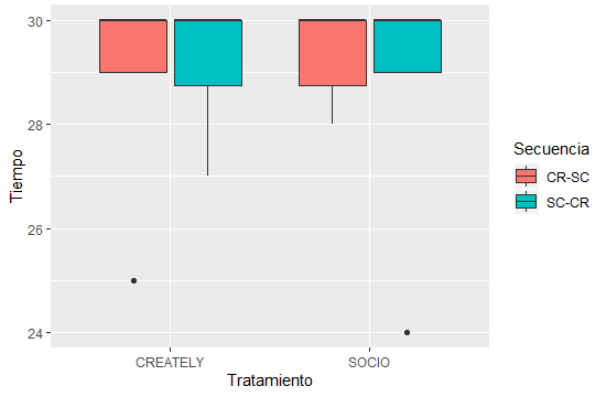


Figura D.3: *Boxplot* del tiempo utilizado para realizar la tarea según tratamiento-secuencia.

La Figura D.4 presenta el *boxplot* asociado al tiempo de realización de la tarea según tratamiento-periodo. Al igual que en el caso anterior, no parece haber diferencias estadísticamente significativas entre los dos tratamientos en ambos periodos.

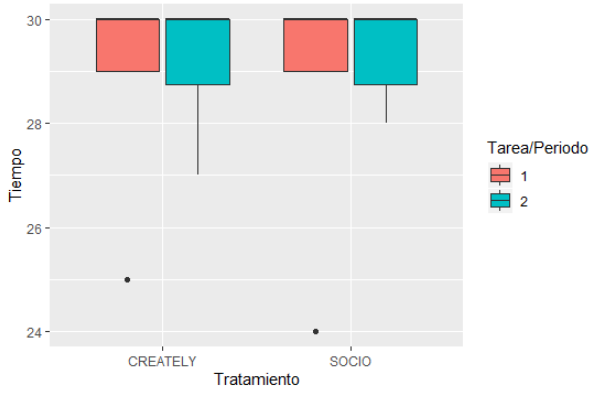


Figura D.4: *Boxplot* del tiempo empleado en realizar la tarea según tratamiento-periodo.

Número de mensajes de discusión

En la Figura D.5 se representa el *boxplot* del nº de mensajes intercambiados de discusión entre los integrantes de cada equipo para realizar las tareas según tratamiento-secuencia. Como se muestra en esta figura, la secuencia CR-SC genera un mayor nº de mensajes para Creately en relación con SOCIO, pero en la secuencia SC-CR hay más intercambio de mensajes para SOCIO. También se puede ver que con SOCIO se genera un nº similar de mensajes para ambas secuencias, aunque ligeramente superior para SC-CR. No obstante, con Creately se generan más mensajes en CR-SC respecto a SC-CR.

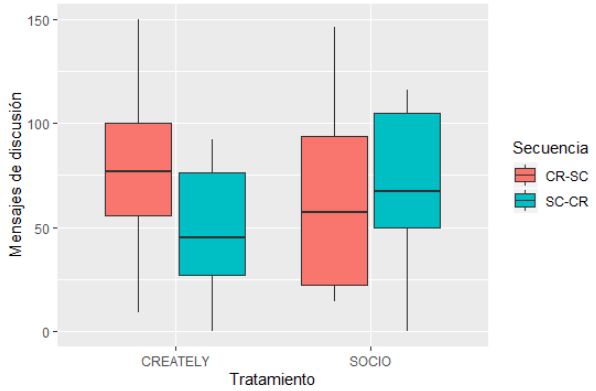


Figura D.5: *Boxplot* del nº de mensajes de discusión según tratamiento-secuencia.

En la Figura D.6 se presenta el *boxplot* asociado al nº de mensajes intercambiados de discusión según tratamiento-tarea. Como se observa, en el caso de Creately se produce un mayor nº de mensajes intercambiados de discusión para la tarea 1 que para la tarea 2, mientras que en el caso de SOCIO se genera un nº de mensajes similar para ambas tareas, aunque ligeramente superior para la primera. Comparando estos resultados, observamos que en la tarea 1 se genera un mayor intercambio de mensajes para Creately respecto a SOCIO; en cambio, en la tarea 2 ocurre lo contrario.

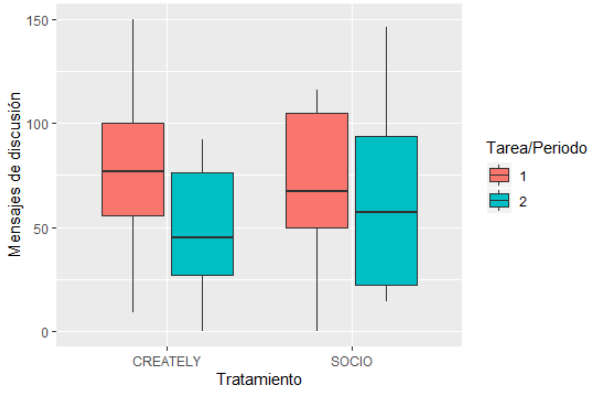


Figura D.6: *Boxplot* del nº de mensajes de discusión según tratamiento-periodo.

D.3. Satisfacción

En la Figura D.7 se muestra el *boxplot* de las valoraciones de satisfacción de los sujetos según tratamiento-secuencia. En relación con CR-SC, Creately presenta superiores valoraciones; por el contrario, en SC-CR, las valoraciones para ambos tratamientos se encuentran muy equiparadas.

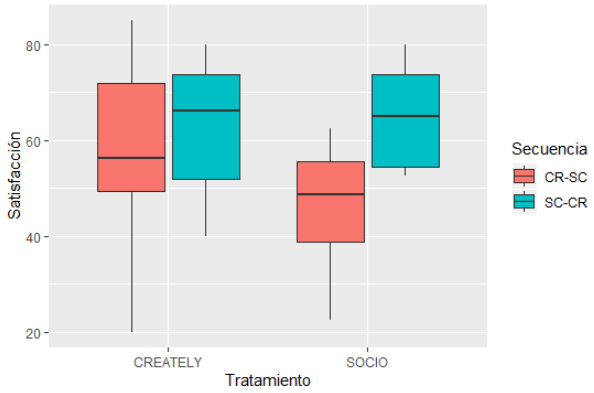


Figura D.7: *Boxplot* de las valoraciones de satisfacción según tratamiento-secuencia.

En la Figura D.8 se muestra el *boxplot* de las valoraciones de satisfacción según tratamiento-periodo. Como podemos ver, Creately obtiene una puntuación mayor de satisfacción en la tarea 2, lo cual coincide con lo observado en la Figura D.7 (Creately presenta mejores resultados en SC-CR). Además, el chatbot SOCIO obtiene superiores valoraciones en la realización de la tarea 1.

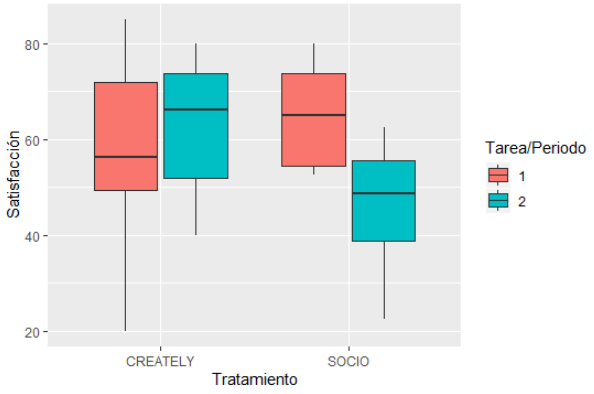


Figura D.8: *Boxplot* de las valoraciones de satisfacción según tratamiento-periodo.

D.4. Calidad

Las métricas empleadas para evaluar la calidad de los diagramas obtenidos son: la accuracy, la precisión, el recall, los aciertos y el error. Seguidamente, se muestran sus *boxplots* según tratamiento-secuencia, y según tratamiento-tarea.

Accuracy

El *boxplot* de las valoraciones de accuracy de los diagramas creados por los sujetos mediante las herramientas SOCIO y Creately según tratamiento-secuencia se muestra en la Figura D.9. Los diagramas creados a través de Creately presentan mejores valoraciones de accuracy para CR-SC respecto a SC-CR, en cambio los diagramas creados mediante SOCIO obtienen una valoración mayor en SC-CR respecto a CR-SC. Al realizar la comparativa entre ambos tratamientos, Creately presenta una puntuación mayor que SOCIO en CR-SC; por lo contrario, en SC-CR se obtiene una puntuación mayor con SOCIO.

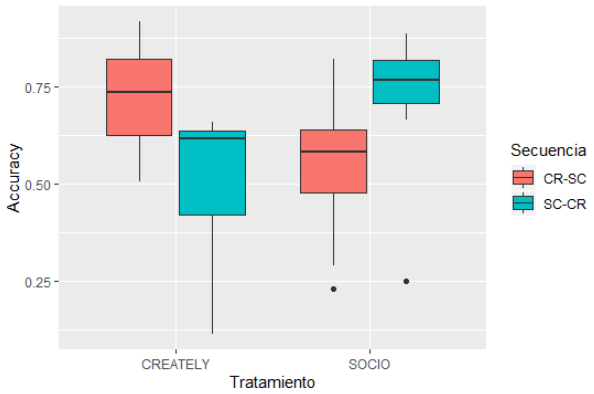


Figura D.9: *Boxplot* de las valoraciones de accuracy según tratamiento-secuencia.

En la Figura D.10 se observa el *boxplot* de las valoraciones de accuracy de los diagramas creados por los sujetos experimentales con SOCIO y Creately según tratamiento-periodo. Como podemos ver,

ambos tratamientos obtienen mejores resultados para la tarea 1. Además, en esta tarea, las valoraciones para SOCIO son mejores que para Creately, mientras que en el caso de la tarea 2 las valoraciones son similares para ambos tratamientos.

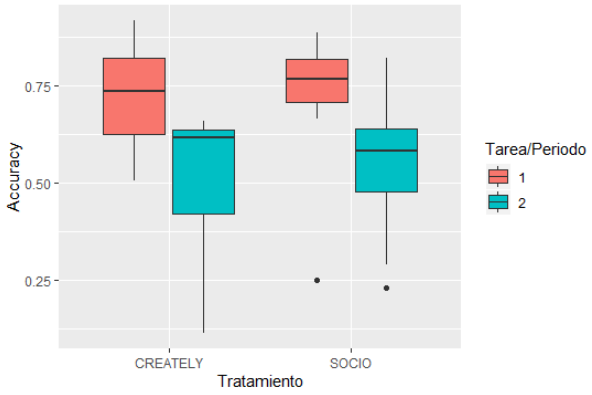


Figura D.10: *Boxplot* de las valoraciones de accuracy según tratamiento-periodo.

Precisión

En la Figura D.11 se muestra el *boxplot* de las valoraciones de precisión de los diagramas creados por los sujetos utilizando SOCIO y Creately según tratamiento-secuencia. Tal como se puede comprobar, en CR-SC, Creately presenta mejores resultados, mientras que en SC-CR se obtiene una mayor puntuación con SOCIO. Además, Creately obtiene mayor puntuación en CR-SC respecto a SC-CR; por el contratio, SOCIO obtiene mayor puntuación en SC-CR respecto a CR-SC.

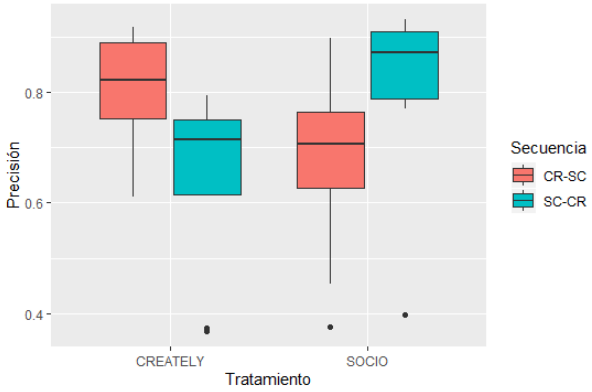


Figura D.11: *Boxplot* de las valoraciones de precisión según tratamiento-secuencia.

En la Figura D.12 se presenta el *boxplot* de las valoraciones de precisión de los diagramas creados por los sujetos a través de SOCIO y Creately según tratamiento-periodo. Se comprueba que para la tarea 1 se ha obtenido una puntuación mayor de precisión con SOCIO, mientras que en la tarea 2 se ha obtenido una puntuación similar para ambas herramientas. Comparando ambos tratamientos, SOCIO y Creately presentan valoraciones superiores en la tarea 1 respecto a la tarea 2.

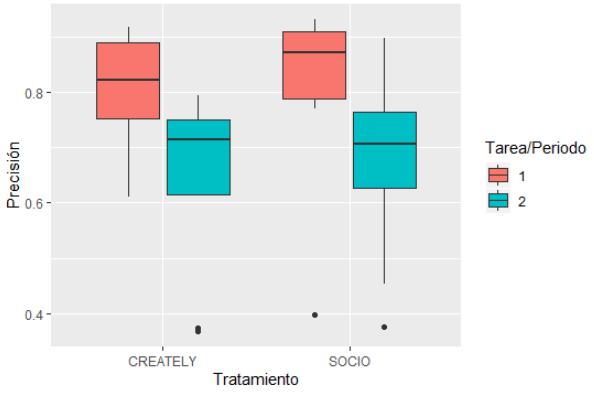


Figura D.12: *Boxplot* de las valoraciones de precisión según tratamiento-periodo.

Recall

En la Figura D.13 se presenta el *boxplot* de las valoraciones de recall de los diagramas creados por los sujetos utilizando SOCIO y Creately según tratamiento-secuencia. Los creados con Creately presentan valoraciones superiores en CR-SC, aunque los diagramas creados con SOCIO tienen mejores valoraciones en SC-CR. Comparando ambos tratamientos, podemos observar que en la secuencia CR-SC se obtienen mejores puntuaciones de recall con Creately respecto a SOCIO; por el contrario, en la secuencia SC-CR se obtienen mejores puntuaciones con SOCIO.

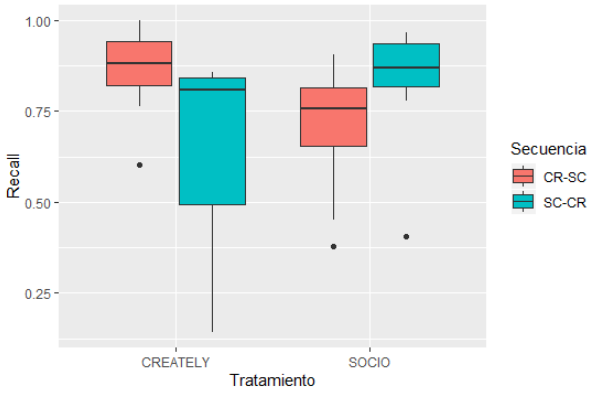


Figura D.13: *Boxplot* de las valoraciones de recall según tratamiento-secuencia.

En la Figura D.14 se muestra el *boxplot* en relación con las valoraciones de recall de los diagramas creados por los sujetos utilizando SOCIO y Creately según tratamiento-tarea. Ambas herramientas, SOCIO y Creately presentan mejores valoraciones en tarea 1 respecto a tarea 2. Si se comparan los tratamientos, las valoraciones para la tarea 1 son idénticas para SOCIO y Creately; sin embargo, en el caso de la tarea 2 son algo superiores para Creately, aunque también son mucho más dispersas inferiormente.

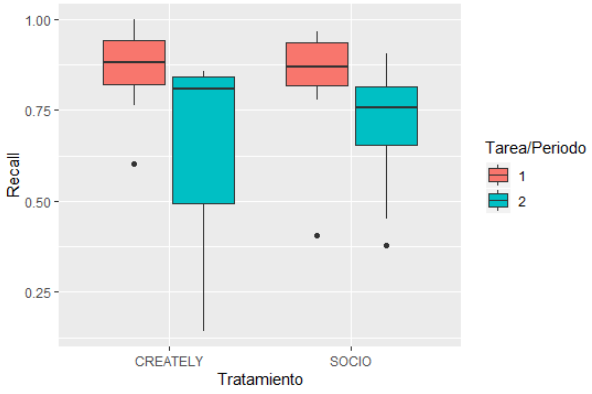


Figura D.14: *Boxplot* de las valoraciones de recall según tratamiento-periodo.

Aciertos

En la Figura D.15 se muestra el *boxplot* en relación con las valoraciones de la métrica de aciertos de los diagramas creados por los sujetos mediante las herramientas SOCIO y Creately según tratamiento-secuencia. A través de Creately se han creado diagramas que presentan valoraciones superiores de aciertos para CR-SC; por el contrario, en el caso de SOCIO se obtienen mejores valoraciones para SC-CR. Realizando la comparativa entre los dos tratamientos, las valoraciones son superiores en Creately en relación con las valoraciones de SOCIO para CR-SC; en cambio, SOCIO presenta valoraciones mejores para SC-CR.

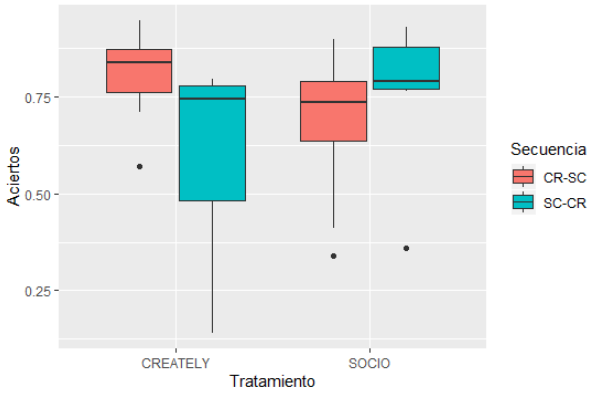


Figura D.15: *Boxplot* de las valoraciones de aciertos según tratamiento-secuencia.

En la Figura D.16 se presenta el *boxplot* de las valoraciones de aciertos de los diagramas creados por los sujetos a través de las herramientas SOCIO y Creately según tratamiento-periodo. Como podemos ver, tanto mediante SOCIO como a través de Creately se obtienen valoraciones superiores en la tarea 1 respecto a la tarea 2. Por una parte, considerando las dos herramientas, Creately obtiene valoraciones algo mayores en la tarea 1 que SOCIO. Por otra parte, en la tarea 2 ambas obtienen una puntuación similar, aunque en el caso de Creately son más dispersas inferiormente.

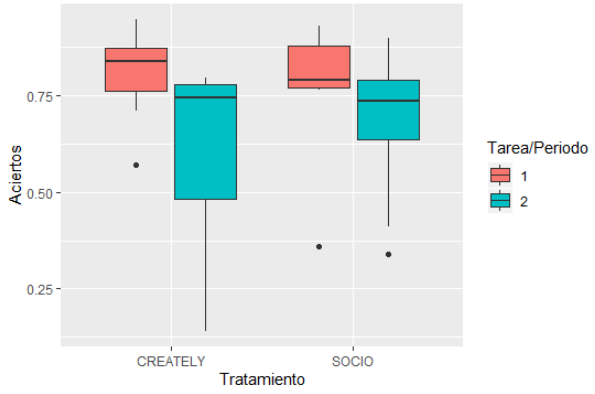


Figura D.16: *Boxplot* de las valoraciones de aciertos según tratamiento-periodo.

Error

En la Figura D.17 se muestra el *boxplot* de las valoraciones de errores perpetrados en los diagramas creados por los sujetos utilizando SOCIO y Creately según tratamiento-secuencia.

Como podemos observar, la aplicación web Creately muestra errores mayores en SC-CR; por el contrario, en el caso de SOCIO se obtienen mayores valoraciones de error en CR-SC. Comparando ambas herramientas, con respecto a CR-SC se observa una puntuación mayor de errores con SOCIO; por el contrario, en SC-CR la herramienta Creately es la que presenta un valor mayor de errores.

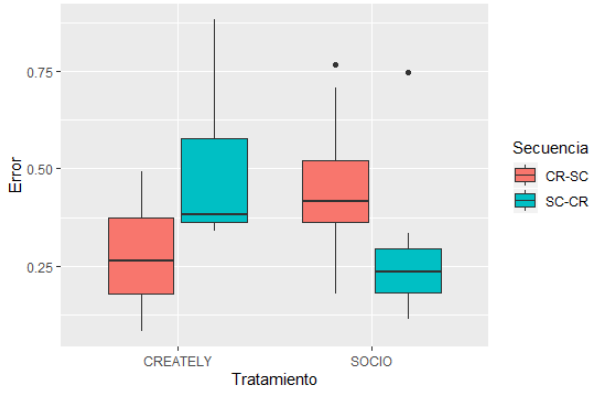


Figura D.17: *Boxplot* de las valoraciones de error según tratamiento-secuencia.

En la Figura D.18 se presenta el *boxplot* de las valoraciones de error de los diagramas creados por los sujetos utilizando SOCIO y Creately según tratamiento-periodo. La evidencia muestra que las dos herramientas obtienen mayores valoraciones de la métrica de error en la tarea 2, es decir, en ella los resultados son peores.

Tanto en la tarea 1 como en la 2 se obtienen puntuaciones similares para SOCIO y Creately, aunque para Creately son más dispersas.

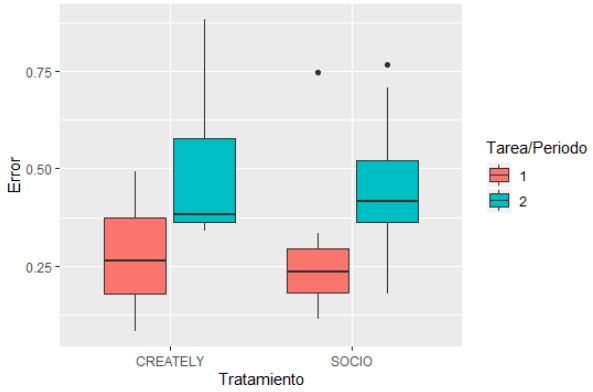


Figura D.18: *Boxplot* de las valoraciones de error según tratamiento-periodo.

E

Datos particulares de SOCIO

Las tareas realizadas a través del chatbot SOCIO han dado lugar a un mayor número de datos para analizar, estos son nº de mensajes: útiles, dirigidos a SOCIO, de error por fallos cometidos por los usuarios, descriptivos, correctos interpretados erróneamente por SOCIO, y de errores totales. Además, del nº de acciones y comandos.

En la sección E.1 se examinan los datos mencionados. Para llevar a cabo este análisis, se ha empleado el lenguaje R para la obtención del objeto de estudio [5]. En la sección E.2 se discuten las conclusiones obtenidas.

E.1. Análisis de los datos

Inicialmente, un análisis gráfico mediante *boxplots* ha sido realizado. Dichos *boxplots* representan la información correspondiente a cada una de las métricas agrupada por tarea. Posteriormente, se ha llevado a cabo, para cada una de las métricas a estudiar, un t-test para muestras independientes, según [5], en el que se confrontan las medias correspondientes a las tareas 1 y 2. Finalmente, se calcula el tamaño de efecto [6] de cada una de las tareas a través de la d de Cohen y su correspondiente error estándar (SE), según [1].

E.1.1. Mensajes enviados a SOCIO

El nº de mensajes dirigidos al chatbot por parte de un sujeto experimental a lo largo de una tarea abarca también al nº de mensajes de error. Se ha diferenciado entre los mensajes erróneos por fallos cometidos por los usuarios y los mensajes correctos interpretados erróneamente por el chatbot SOCIO. Los mensajes erróneos son aquellos que por fallos en la escritura no son dirigidos al bot a pesar de que el propósito inicial es que éste los recibiera, o bien, aquellos que llegan a enviarse pero no son entendidos por el bot.

Número de mensajes enviados a SOCIO

En la Figura E.1 se representa el *boxplot* relativo al nº de mensajes enviados a SOCIO. Como podemos ver, este número es mucho mayor en la tarea 2 que en la tarea 1.

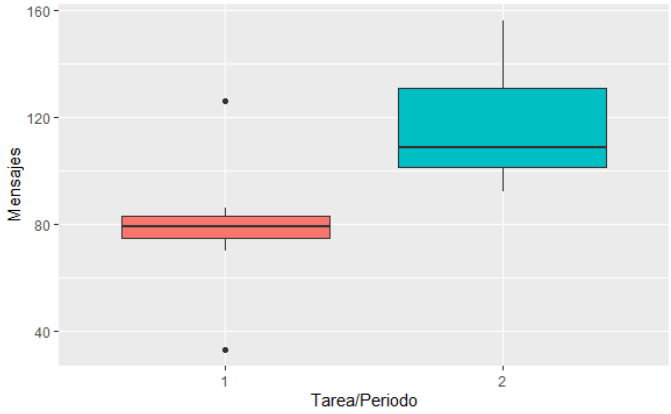


Figura E.1: *Boxplot* del nº de mensajes enviados al chatbot SOCIO.

La Tabla E.1 muestra las conclusiones obtenidas del t-test, al comparar el promedio del nº de mensajes enviados a SOCIO en cada una de las tareas.

Task 1	Task 2	95 %IC	p-value
78.9	117.8	[-65.41, -12.34]	**0.007**

Tabla E.1: Media del nº de mensajes dirigidos al chatbot.

Como podemos ver en la Tabla E.1, la media del nº de mensajes enviados a SOCIO durante la segunda de las tareas es superior con respecto a la primera. De hecho, dicha discrepancia es significativa, lo cual se debe a que p-valor = 0.007, con un intervalo de confianza (IC) del 95 % = [-65.41, -12.34]. Por último, se obtiene que el tamaño del efecto es grande, ya que d = -1.57, SE(d) = 0.33. En definitiva, **el nº de mensajes enviados al chatbot SOCIO en la segunda de las tareas es superior con respecto a los enviados en la primera.**

Número de mensajes erróneos por fallos de los equipos

En la Figura E.2 se representa el *boxplot* relativo al nº de mensajes de error debidos a fallos cometidos por los equipos a lo largo de la elaboración de las tareas. Como se puede observar, dicha cantidad es superior en la segunda tarea que en la primera.

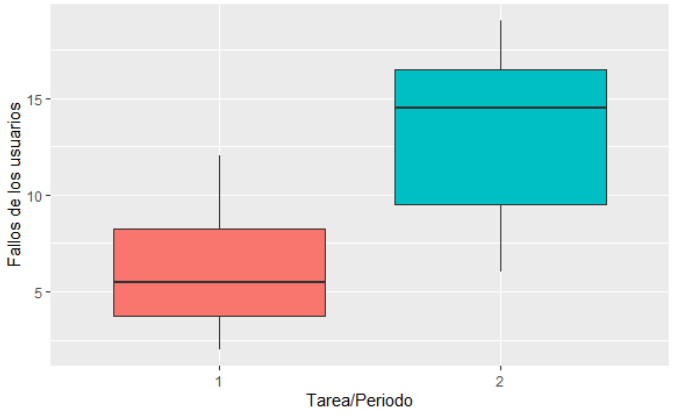


Figura E.2: *Boxplot* del nº de mensajes erróneos cometidos por los equipos.

La Tabla E.2 muestra las conclusiones obtenidas del t-test, al comparar el promedio del nº de mensajes erróneos por fallos de los equipos en cada una de las tareas.

Task 1	Task 2	95 %IC	p-value
6.13	13.3	[-11.58, -2.67]	**0.004**

Tabla E.2: Media del nº de mensajes erróneos debidos a los equipos.

Como podemos ver en la Tabla E.2, la media del nº de mensajes erróneos por fallos cometidos por los equipos es superior en la segunda tarea que en la primera. Dicha discrepancia es significativa estadísticamente, lo cual se debe a que p-valor = 0.004, con un IC del 95 % = [-11.58, -2.67]. Por último, se obtiene que el tamaño del efecto es grande, ya que d = -1.73, SE(d) = 0.34.

En definitiva, **el nº de mensajes erróneos por fallos cometidos por los equipos originados en la segunda tarea es superior con respecto a los originados durante la primera.**

Número de mensajes correctos interpretados erróneamente por SOCIO

En la Figura E.3 se representa el *boxplot* relativo al nº de mensajes correctos interpretados erróneamente por SOCIO a lo largo de la elaboración de las tareas. Como podemos ver, el nº de dichos errores es superior en la segunda tarea que en la primera.

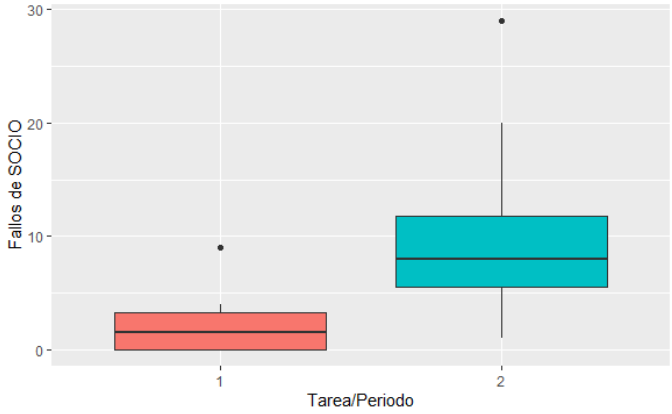


Figura E.3: *Boxplot* del nº de mensajes correctos interpretados erróneamente por el chatbot SOCIO.

La Tabla E.3 muestra las conclusiones obtenidas del t-test, al comparar el promedio del nº de mensajes correctos interpretados erróneamente por el chatbot SOCIO en ambas tareas.

Task 1	Task 2	95 %IC	p-value
2.37	10.37	[-16.13, 0.13]	0.053

Tabla E.3: Media del nº de mensajes correctos interpretados erróneamente por el chatbot.

Como podemos ver en la Tabla E.3, el promedio del nº mensajes correctos interpretados erróneamente por SOCIO es superior en la segunda tarea que en la primera. No obstante, esta discrepancia no es estadísticamente significativa, lo cual se debe a que p-valor = 0.053, con un IC del 95 % = [-16.13, 0.13]. Por último, se obtiene que el tamaño del efecto es grande, ya que d = -1.12, SE(d) = 0.29.

En definitiva, aunque los resultados obtenidos no presentan diferencias significativas, **el nº de mensajes correctos interpretados erróneamente por SOCIO es superior en la segunda tarea que en la primera.**

Número de mensajes erróneos enviados a SOCIO

El *boxplot* asociado al nº de mensajes de error enviados a SOCIO a lo largo de la elaboración de las tareas se muestra en la Figura E.4. Como podemos ver, el nº de mensajes de error originados en la segunda tarea es superior con respecto a los de la primera.

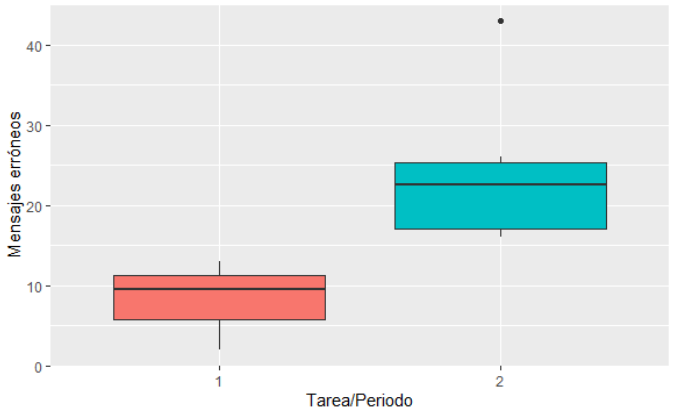


Figura E.4: *Boxplot* del nº de mensajes erróneos dirigidos al chatbot SOCIO.

La Tabla E.4 muestra las conclusiones obtenidas del t-test, al comparar los promedios del nº de mensajes de error enviados al chatbot durante ambas tareas.

Task 1	Task 2	95 %IC	p-value
8.5	23.62	[-22.74, -7.5]	**0.0013**

Tabla E.4: Media del nº de mensajes de error dirigidos a SOCIO.

Como podemos ver en la Tabla E.4, el promedio de mensajes de error es superior en la segunda tarea que en la primera. Dicha discrepancia es significativa estadísticamente, lo cual se debe a que p-valor = 0.0013, con un IC del 95 % = [-22.74, -7.5]. Por último, se obtiene que el tamaño del efecto es grande, ya que d = -2.23, SE(d) = 0.41.

En definitiva, **el nº de mensajes erróneos originados en la segunda tarea es considerablemente superior que los ocasionados en la primera tarea.**

E.1.2. Mensajes útiles enviados a SOCIO

Dentro del nº de mensajes útiles dirigidos al chatbot se engoba a todos los que han contribuido en la elaboración del diagrama desarrollado por los sujetos experimentales a lo largo de las tareas.

Hay dos tipos de mensajes útiles: comandos (como por ejemplo, */talk remove home*) y mensajes descriptivos (como, por ejemplo, */talk the room contains windows and the house includes doors*).

Número de mensajes útiles enviados a SOCIO

El *boxplot* relativo al nº de mensajes útiles enviados a SOCIO a lo largo del desarrollo de las tareas se muestra en la Figura E.5, esto es, los mensajes que contribuyeron al diagrama desarrollado por los sujetos. Como podemos ver, el nº de mensajes útiles es superior en la segunda tarea con respecto a la primera.

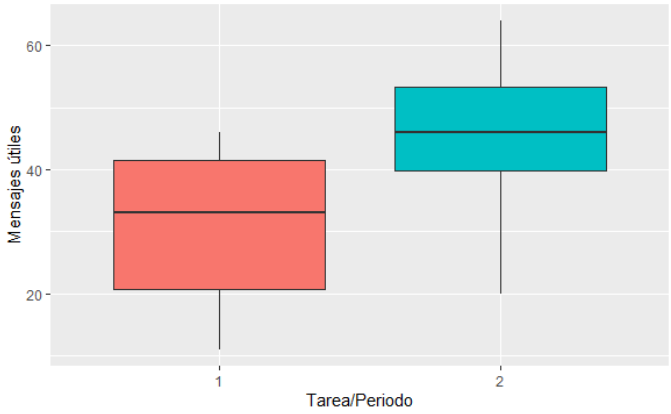


Figura E.5: *Boxplot* del nº de mensajes útiles dirigidos a SOCIO.

La Tabla E.5 muestra las conclusiones obtenidas del t-test, al comparar los promedios del nº de mensajes útiles enviados al chatbot durante las tareas.

Task 1	Task 2	95 %IC	p-value
31	44.25	[-27.8, 1.3]	0.07

Tabla E.5: Media del nº de mensajes útiles dirigidos a SOCIO.

Como podemos ver en la Tabla E.5, el promedio del nº mensajes útiles enviados al chatbot SOCIO es superior en la segunda tarea con respecto a la primera. No obstante, esta discrepancia no es significativa estadísticamente, lo cual se debe a que p-valor = 0.07, con un IC del 95 % = [-27.8, 1.3]. Por último, se obtiene que el tamaño del efecto es grande, ya que d = -0.97, SE(d) = 0.28.

En definitiva, a pesar de que los resultados obtenidos no son significativos, **el nº de mensajes útiles originados durante la segunda tarea es superior con respecto a los originados durante la primera.**

Número de mensajes descriptivos

El *boxplot* relativo al nº de mensajes descriptivos enviados al chatbot SOCIO que han contribuido a la elaboración del diagrama desarrollado por los sujetos se muestra en la Figura E.6. Como podemos ver, el nº de mensajes descriptivos en la segunda tarea es superior que en la primera.

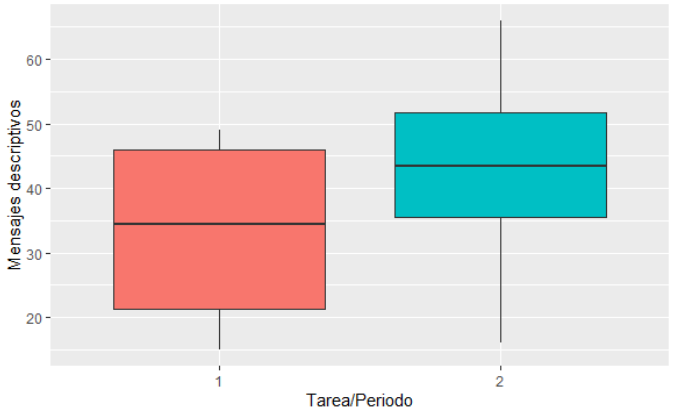


Figura E.6: *Boxplot* del nº de mensajes descriptivos dirigidos a SOCIO.

La Tabla E.6 muestra las conclusiones obtenidas del t-test, al comparar los promedios del nº de mensajes descriptivos enviados al chatbot durante ambas tareas.

Task 1	Task 2	95 %IC	p-value
33	43.5	[-26.76, 5.76]	0.18

Tabla E.6: Media del nº de mensajes descriptivos enviados al chatbot.

Como podemos ver en la Tabla E.6, el promedio del nº de mensajes descriptivos dirigidos al chatbot SOCIO durante la segunda tarea es superior que los enviados en la primera tarea. La discrepancia no es significativa estadísticamente, lo cual se debe a que p-valor = 0.18, con un IC del 95 % = [-26.76, 5.76]. Por último, se obtiene que el tamaño del efecto es mediano, ya que d = -0.69, SE(d) = 0.26.

En definitiva, aunque los resultados obtenidos no son estadísticamente significativos, **el nº de mensajes descriptivos originados durante la segunda tarea es superior que los originados en la primera.**

Número de comandos

En la Figura E.7 se representa el *boxplot* relativo al nº de comandos enviados al chatbot en el desarrollo del diagrama elaborado por los sujetos. Como podemos ver, el nº de comandos enviados durante la segunda tarea es superior que los enviados en la primera.

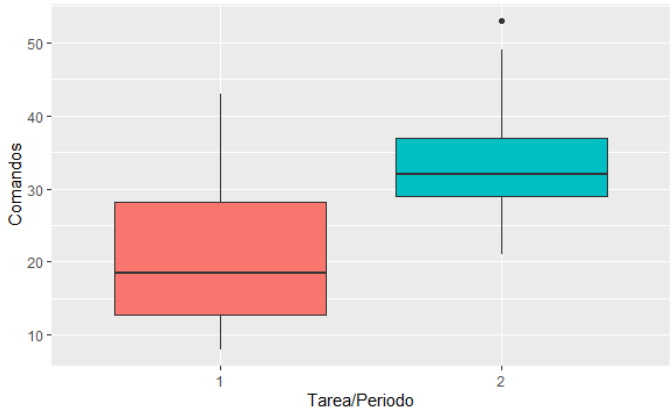


Figura E.7: *Boxplot* del nº de comandos dirigidos al chatbot SOCIO.

La Tabla E.7 muestra las conclusiones obtenidas del t-test, al comparar los promedios del nº de comandos dirigidos al chatbot durante ambas tareas.

Task 1	Task 2	95 %IC	p-value
22	34.5	[-25.93, 0.43]	0.057

Tabla E.7: Media del nº de comandos dirigidos al chatbot.

Como podemos ver en la Tabla E.7, el promedio del nº de comandos dirigidos a SOCIO en la segunda tarea es superior que en la primera. No obstante, esta discrepancia no es significativa estadísticamente, lo cual se debe a que p-valor = 0.057, con un IC del 95 % = [-25.93, 0.43]. Por último, se obtiene que el tamaño del efecto es grande, ya que d = -1.04, SE(d) = 0.28.

En definitiva, aunque los resultados obtenidos no son estadísticamente significativos, **el nº de comandos dirigidos al chatbot en la segunda tarea es superior que los enviados en la primera.**

E.1.3. Acciones desencadenadas

Cuando SOCIO recibe los mensajes para realizar el diagrama de clases debe interpretarlos y genera tres tipos de acciones: crear, modificar o eliminar un elemento. En la Figura E.8 se representa el *boxplot* relativo al nº de acciones desencadenadas por el chatbot a lo largo del desarrollo del diagrama. Como podemos ver, el nº de acciones desencadenadas por SOCIO en la segunda tarea es superior que las de la primera.

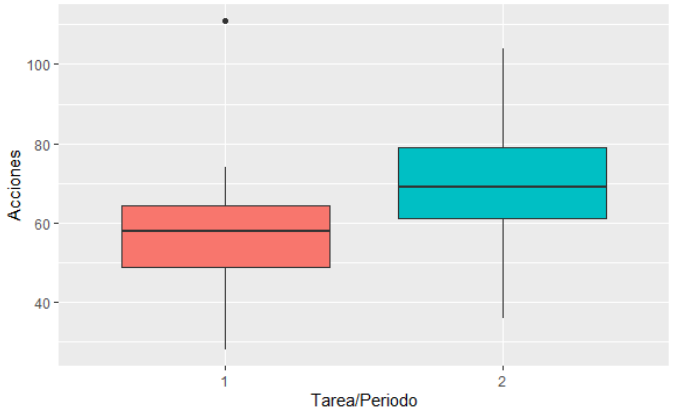


Figura E.8: *Boxplot* del nº de acciones desencadenadas por el chatbot SOCIO.

La Tabla E.8 muestra las conclusiones obtenidas del t-test, al comparar los promedios del nº de acciones desencadenadas por el bot durante ambas tareas.

Task 1	Task 2	95 %IC	p-value
60.87	69.87	[-33.53, 15.53]	0.44

Tabla E.8: Media del nº de acciones desencadenadas por el chatbot.

Como podemos ver en la Tabla E.8, el promedio del nº de acciones desencadenadas durante la segunda tarea es superior que las desencadenadas en la primera tarea. Dicha discrepancia no es significativa estadísticamente, lo cual se debe a que p-valor = 0.44, con un IC del 95 % = [-33.53, 15.53]. Por último, se obtiene que el tamaño del efecto es pequeño, ya que d = -0.39, SE(d) = 0.25.

En definitiva, aunque los resultados obtenidos no son estadísticamente significativos, **el nº de acciones desencadenadas por el chatbot en la segunda tarea es superior que las desencadenadas durante la primera tarea.**

E.2. Discusión de los resultados

La Tabla E.9 muestra las conclusiones obtenidas de los t-test realizados a cada una de las métricas con el fin de comparar las discrepancias observadas entre ambas tareas utilizando el chatbot SOCIO. En la columna t-test, el símbolo + indica que hay diferencias significativas, mientras que el símbolo / indica que no las hay.

Métricas	Hipótesis	t-test	Tamaño del efecto
Mensajes dirigidos a SOCIO	H.S.1.0	+	Grande
Mensajes erróneos por fallos cometidos por los equipos	H.S.2.0	+	Grande
Mensajes correctos interpretados erróneamente por SOCIO	H.S.3.0	/	Grande
Mensajes erróneos enviados a SOCIO	H.S.4.0	+	Grande
Mensajes útiles dirigidos a SOCIO	H.S.5.0	/	Grande
Mensajes descriptivos enviados a SOCIO	H.S.6.0	/	Mediano
Comandos enviados a SOCIO	H.S.7.0	/	Grande
Acciones desencadenadas por SOCIO	H.S.8.0	/	Pequeño

Tabla E.9: Resumen de los resultados experimentales tras la interacción con el chatbot SOCIO.

Como se puede observar, no hay discrepancias significativas causadas por la tarea, salvo para las métricas de los mensajes enviados a SOCIO, los mensajes erróneos por fallos cometidos por los equipos y los mensajes erróneos enviados a SOCIO. Por lo tanto, las hipótesis H.S.3.0, H.S.5.0, H.S.6.0, H.S.7.0 y H.S.8.0 no pueden ser rechazadas. A pesar de que en dichos casos las diferencias no son significativas estadísticamente (lo que puede estar provocado por un pequeño tamaño muestral), atendiendo al tamaño del efecto aparentemente se genera un nº de mensajes correctos interpretados erróneamente por SOCIO, de mensajes útiles, descriptivos y de comandos superior en la segunda tarea, de modo que parece que la segunda tarea requiere mayor esfuerzo que la primera.

Sin embargo, para las métricas de los mensajes enviados a SOCIO, los mensajes erróneos por fallos cometidos por los equipos y los mensajes erróneos enviados a SOCIO, las diferencias en las tareas son estadísticamente significativas, y además, son grandes. Por lo tanto, podemos rechazar las hipótesis H.S.1.0, H.S.2.0 y H.S.4.0.

F

Gráficos cuantil-cuantil

En este apéndice se muestran los gráficos cuantil-cuantil obtenidos para observar cómo de cerca se encuentra la distribución de los datos de la familia de experimentos con respecto a la distribución normal. Para ello, hemos desarrollado un gráfico de probabilidad normal para cada una de las métricas. A mayor cercanía entre la nube de puntos y la recta en cada uno de los gráficos, más se aproxima la distribución de los datos a una distribución normal.

En las Figuras F.1-F.9 se muestra el gráfico correspondiente para cada una de las métricas. Como se puede observar, en ninguno de los gráficos hay diferencias relevantes entre la nube de puntos y la recta. En particular, en la parte central de los gráficos (que es donde se encuentran la mayoría de los datos) ambos elementos (la recta y la nube de puntos) son muy parecidos.

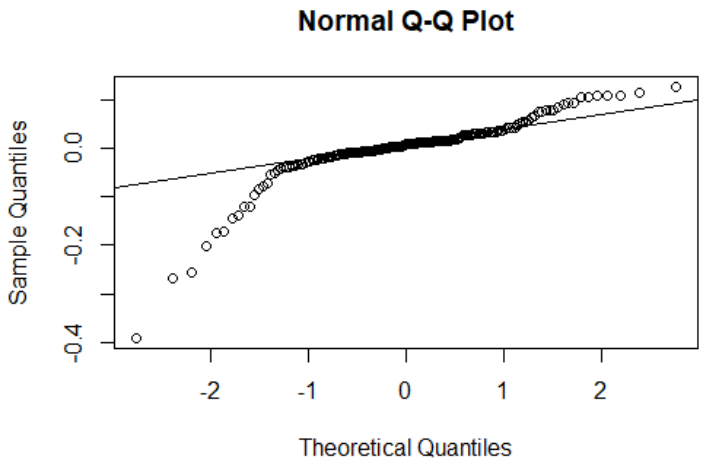


Figura F.1: Representación de probabilidad normal de la completitud.

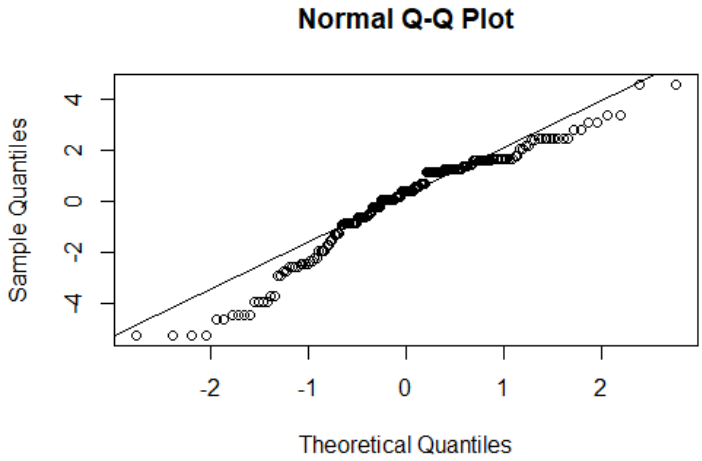


Figura F.2: Representación de probabilidad normal del tiempo empleado en realizar una tarea.

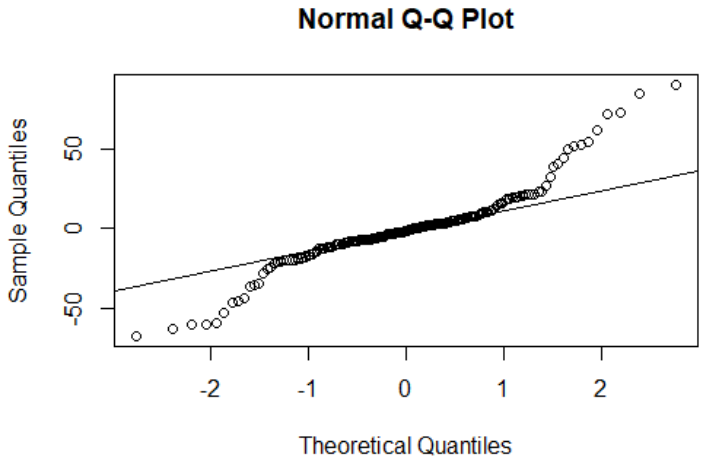


Figura F.3: Representación de probabilidad normal del número de mensajes de discusión.

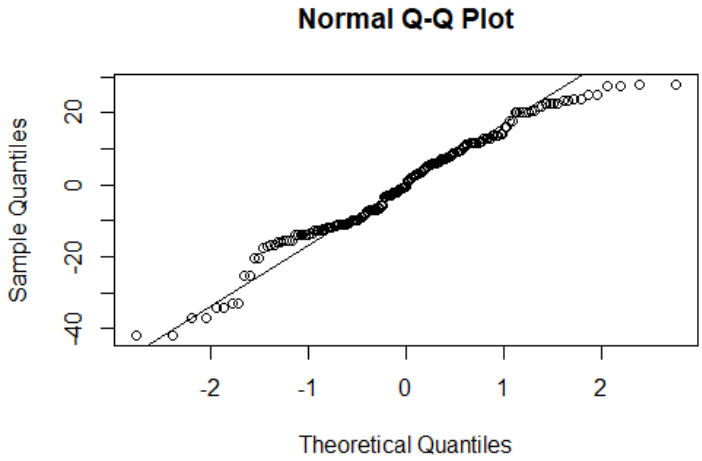


Figura F.4: Representación de probabilidad normal de la satisfacción.

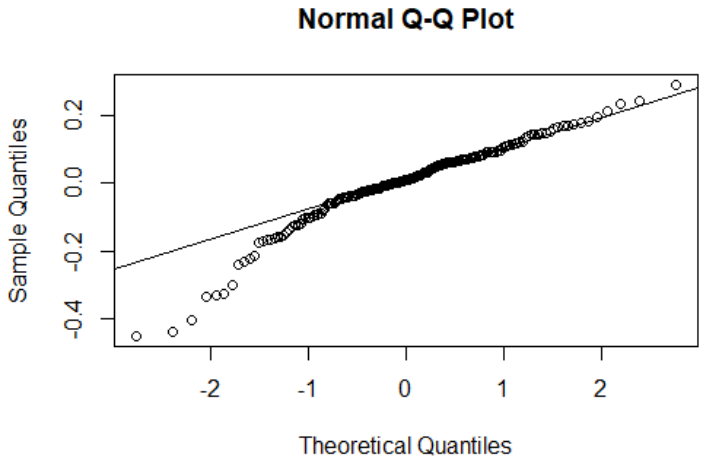


Figura F.5: Representación de probabilidad normal de la métrica accuracy.

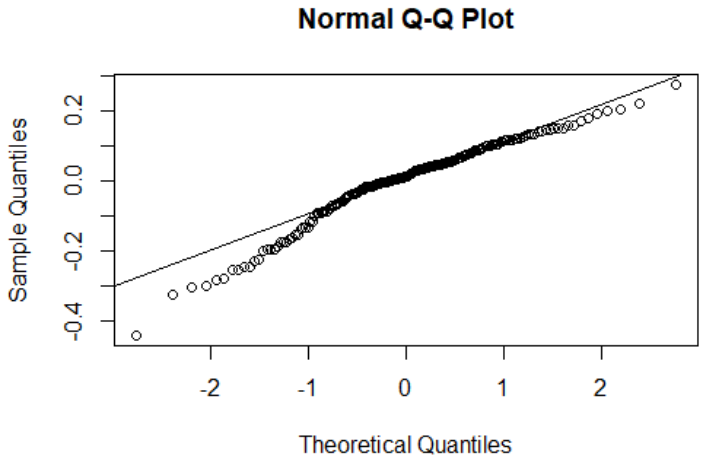


Figura F.6: Representación de probabilidad normal de la métrica precisión.

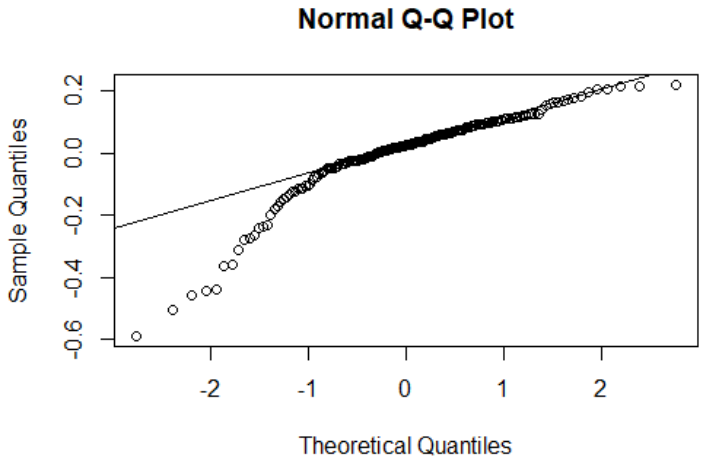


Figura F.7: Representación de probabilidad normal de la métrica recall.

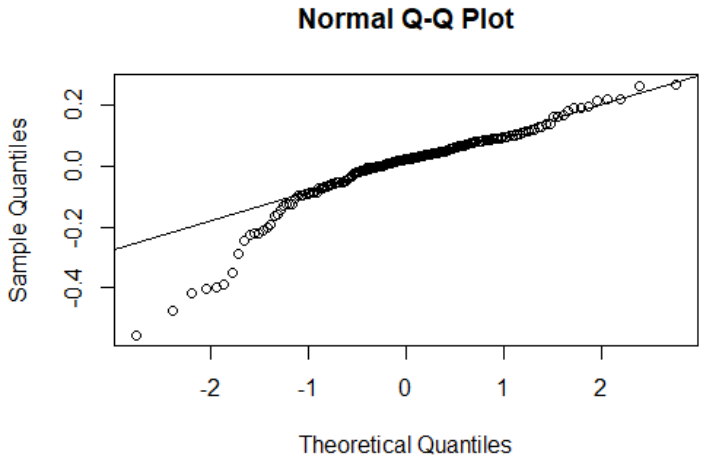


Figura F.8: Representación de probabilidad normal de la métrica aciertos.

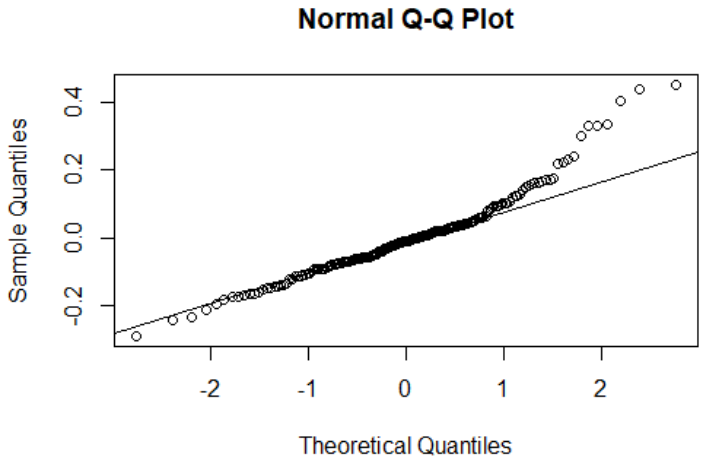


Figura F.9: Representación de probabilidad normal de la métrica error.


UAM
Universidad Autónoma
de Madrid